

Chapter 2

The Causal Metric Hypothesis

Abstract. This chapter formally introduces the *causal metric hypothesis*, and describes in detail its motivations and justifications. Foremost among these are the *metric recovery theorems* of Hawking and Malament, which state, roughly, that “*the causal structure of relativistic spacetime determines its metric structure up to scale.*” As understood in causal set theory, the novel assumption that spacetime is discrete provides a natural notion of scale, given by the “sizes of fundamental elements and relations.” This suggests that causal structure alone can account for emergent geometry in the discrete context. Section 2.1 introduces a general version of the causal metric hypothesis, which states, very broadly, that “*the properties of the physical universe are manifestations of causal structure.*” This basic idea may be modified and/or interpreted in various ways; in particular, the *strong interpretation* of the causal metric hypothesis ascribes *all* of physics, including “nongravitational matter,” to causal structure at the fundamental scale. Section 2.2 introduces a *classical version* of the causal metric hypothesis, which states that classical spacetime may be modeled in terms of mathematical objects called *directed sets*, or, more conventionally, *directed graphs*. The term “directed set” has a different conventional meaning, but I prefer to re-purpose the term than to use awkward graph-theoretic terminology. Section 2.3 begins the study of metric recovery by introducing five types of structure on relativistic spacetime; namely, *metric*, *conformal*, *causal*, *smooth*, and *topological* structure, in decreasing order of detail. Section 2.4 discusses metric structure in the relativistic context, i.e. *pseudo-Riemannian geometry*. Section 2.5 covers conformal structure, which defines “geometry up to scale.” Section 2.6 discusses causal structure, which generalizes the “null cone structure” on Minkowski spacetime. The metric recovery theorems state that “*causal structure determines conformal structure*” under suitable assumptions. Section 2.7 introduces *causality conditions* on relativistic spacetime, which play a technical role in the metric recovery theorems. Section 2.8 gives a formal statement of metric recovery, sketches its proof, and describes how it motivates the causal metric hypothesis in the discrete setting. Section 2.9 explains why continuum-based theories are inherently awkward for modeling fundamental physics. Section 2.10 outlines some of the basic principles underlying the technical developments of subsequent chapters.

2.1 General Version of the Hypothesis

Background and context. Discrete causal theory is founded on a single motivating idea, which I refer to as the *causal metric hypothesis*. While an informal sketch of this hypothesis appears in Section 1.2, a much deeper analysis of its meaning and implications is required to support the developments in the remainder of the book. The choice of terminology for the causal metric hypothesis is my own, but variants of the same basic idea may be distilled from a number of previous sources, stretching back to at least the 1960s. The clearest and most explicit of these is Rafael Sorkin’s structural *ansatz* for causal set theory, “*order plus number equals geometry*.” This statement reflects the conviction that the metric recovery theorems of Hawking [HA76] and Malament [MA77], proven in the late 1970s, suggest subtle new aspects of spacetime structure *beyond the scope of their native relativistic paradigm*, just as the Lorentz invariance of Maxwell’s equations suggested special relativity, despite the fact that these equations were formulated in the Newtonian context. In the case of metric recovery, the suggested new structure is *discrete directed structure*. Causal set theory, which emerged in the early 1980s, was the first sustained theoretical program attempting to describe spacetime in such terms, although a few abortive individual efforts along similar lines enjoy chronological priority. I postpone further discussion of such historical details until Chapter 3, which provides a concise overview of the origins of discrete causal theory before launching into the technical apparatus of directed sets and multidirected sets. The present chapter involves a modest amount of standard technical material, but focuses as much as possible on basic conceptual topics.

The causal metric hypothesis itself does not require any discreteness assumptions, but the metric recovery theorems described in Section 2.8 provide stronger motivation for the hypothesis in the discrete context than in the continuum-based setting.¹ Sorkin’s version of the hypothesis, which applies at the classical level, is *explicitly* discrete, since it invokes *counting*, and therefore requires a notion of local finiteness in order to make sense.² This book devotes the most attention to versions of the hypothesis that are very similar to Sorkin’s. However, it is possible to imagine continuous versions of the causal metric hypothesis, or versions that are neither discrete nor continuous. Further, due to the ubiquity of directed relationships in modern science, there exist many “non-fundamental” settings in which the causal metric hypothesis serves as a useful source of *analogy*, without being taken literally. For example, one may compare certain abstract architectures in computer science to spacetime, and may define “frames of reference,” and other related notions for these structures, without insisting on any exact correspondence between the two, or sug-

¹Indeed, in the latter setting, the metric recovery theorems essentially say that causal structure is *not quite sufficient* to recover geometry, at least under relativistic assumptions.

²The “local finiteness” condition used in causal set theory, which I refer to more descriptively as *interval finiteness* (IF), is not necessarily ideal for this purpose, as explained in Chapter 4. However, causal sets actually appearing in physically realistic scenarios in the literature generally *do* satisfy a suitable notion of local finiteness, even in the context of cosmology.

gesting that spacetime *is* a computer in some sense. In *quantum information theory*, such considerations are more than just an analogy, but this subject is not explored in this book.

Restatement of the general version of the hypothesis. The philosophical content of the causal metric hypothesis is that *the observed properties of the physical universe arise from causal relationships between pairs of events*, or more generally, from causal relationships *among families* of events. The latter generalization is included to allow for the possibility of *classical holism*, although I focus almost exclusively on classically reductionist models in this book. The following statement of the causal metric hypothesis, repeated from Section 1.2, is sufficiently general to use as a starting point:

Definition 2.1.1. Causal metric hypothesis (CMH). *The properties of the physical universe are manifestations of causal structure.*

The causal metric hypothesis may be regarded as an expression of the longstanding idea, examined explicitly by Leibniz, Gauss, Riemann, Einstein, Kaluza and Klein, Weyl, Wheeler, and many others, that physics is *essentially structural* in nature. The hypothesis takes the familiar relationship between cause and effect to be the fundamental building block of this structure.

Scope of the hypothesis; strong interpretation. The proper scope of the causal metric hypothesis is debatable. A conservative approach is to soften the statement in Definition 2.1.1 by replacing the words “physical universe” with the word “spacetime.” This approach abandons any attempt to explain the “material content of spacetime” by means of causal structure. As noted in Section 1.7, this alternative leads to theories that possess, at best, a limited degree of background independence. These include “discrete quantum field theories,” and “theories of gravity,” but not “unified theories” in the deepest sense. At the opposite extreme, one may choose to take the statement in Definition 2.1.1 at face value, and interpret the opening phrase “the properties,” as “*all* the properties.” This is the **strong interpretation** of the causal metric hypothesis. Its radical nature was already highlighted in Section 1.10. The strong interpretation leads to a version of discrete causal theory that is ambitious and optimistic, but also quite pleasing at a structural and aesthetic level. In particular, it enables *perfect background independence*, by removing any possibility of tension between “material bodies” and “spacetime.” The advantages of this approach are elaborated in Section 2.7 in the context of “causality paradoxes,” and are revisited periodically throughout the book. Definition 2.1.1 is deliberately phrased in such a way as to *suggest* the strong interpretation of the causal metric hypothesis, but weaker interpretations are possible, and most of the methods and results of the book do not require the strong interpretation.

Prescription versus description. An important philosophical distinction between the relativistic viewpoint and the causal metric hypothesis, particularly its strong interpretation, involves the choice between *prescription of possible behavior* and *description of actual behavior*. Relativity employs spacetime geometry to *prescribe*

which events *may* influence a given event, while the causal metric hypothesis interprets the same structure as merely an approximate way of *describing* which events *actually do* influence others. This distinction enables discrete causal theory to eliminate “awkward counterfactual speculation” regarding causality in relativity, as described in Section 1.3. In particular, the discrete causal rejoinder to Wheeler’s famous statement that “*spacetime tells matter how to move; matter tells spacetime how to curve*” [WH98], is that “*things happen; “spacetime” and “matter” are ways of describing them.*” A general preference for description over prescription in theoretical physics constitutes one of the philosophical principles informing the overall development of discrete causal theory, as described in Section 2.10. This does not mean that the theory seeks to avoid the necessary criteria of *explaining* and *predicting* physical behavior, but merely that “causal structure” should mean neither more nor less than the aggregate of *actual* causes and effects. This viewpoint is closely linked to the notion of perfect background independence, because it removes the distinction between a “spacetime” that prescribes behavior, and “material participants” in this behavior.

Technical implementations. A “technical implementation” of the causal metric hypothesis specifies what the words *causal structure* in Definition 2.1.1 are taken to mean in mathematical terms. Many different such implementations are possible, both at the classical level and the quantum level. In Section 2.2, I introduce a *classical version* of the causal metric hypothesis (CCMH), which identifies *directed sets* as the chosen mathematical models of classical spacetime. This version is more specific than the version appearing in Definition 2.1.1, but is still quite general, since its purpose is to accommodate any “reasonable” variant of the theory. Hence, additional conditions must be imposed in order to obtain a specific theory capable of precise quantitative description of nature. This is accomplished by specifying a set of *axioms* that restrict attention to a “desirable” class of directed sets, together with a “plausible” physical interpretation of these sets. Physical intuition plays an unavoidable role in this process, but data from experimentally-established physics should be used for guidance whenever possible. Examples of the types of considerations that might be involved in selecting such a set of axioms are the questions of how to implement the idea of *discreteness*, what type of *local behavior* the chosen class of directed sets should exhibit, and whether or not directed sets containing *causal cycles* should be included. Chapters 3 and 4 are largely devoted to identifying a suitable set of axioms for discrete classical causal theory, based both on experimental evidence and on basic structural considerations. A “suggested list” of axioms is offered in Section 4.10, along with “conservative” and “radical” alternatives. A specific choice of axioms leads, via iteration of structure (IS), to a specific version of quantum theory, as described in Part II of the book.

Quantum causal metric hypothesis. A detailed treatment of the *quantum causal metric hypothesis* (QCMH) is postponed until Chapter 7, since much conceptual and technical ground must be covered before it can be adequately explained. In general terms, the role of classical causal structure is superseded in discrete quantum causal theory by the “higher-level multidirected structures” of *kinematic schemes*,

introduced informally in Chapter 1 as “structured configuration spaces” of classical histories, whose “relations” are *co-relative histories*. Given a kinematic scheme, the general strategy is to “superpose” evolutionary processes for its constituent classical histories, thereby building the “quantum universe.” Feynman’s *path summation approach to quantum theory* provides the basic conceptual ingredients of this viewpoint. In a more modern context, this approach shares important features with Isham’s *quantization on a category* [IS05], and Sorkin’s *quantum measure theory* [SO12]. A formal statement of the quantum causal metric hypothesis appears in Section 7.6.

2.2 Classical Version of the Hypothesis

Review of basic building blocks of causal structure. The content of the causal metric hypothesis (CMH) must be expressed mathematically before its physical consequences may be explored in a precise quantitative fashion. In the classical context, this may be accomplished by means of the “classical histories” introduced informally in Chapter 1. In the present chapter, it is necessary to consider the mathematical properties of these histories in slightly more detail, although formal definitions and analysis are postponed until Chapter 3. The basic unit of mathematical structure used to model a *particular* instance of cause and effect between two events is an *ordered pair of abstract elements*, with the first element representing the cause, and the second element representing the effect. In Figure 1.4.1, I introduced a convenient way to represent this structure via diagrams, called *generalized Hasse diagrams*, in which elements are represented by nodes, and relations are represented by directed line segments connecting pairs of nodes, with directions inferred by using an “up the page” convention. For convenience, I reproduce this picture in Figure 2.2.1.

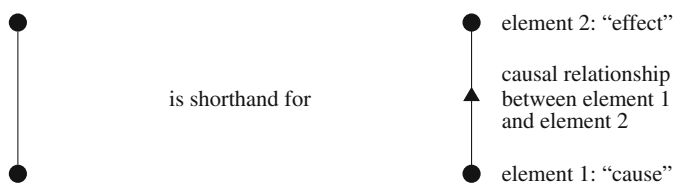


Figure 2.2.1. Abstract representation of a single instance of cause and effect.

Directed sets. Elements 1 and 2 in Figure 2.2.1 are naturally ordered, with element 1, representing the “cause,” preceding element 2, representing the “effect.” It is convenient to name these elements x and y ; one may then write $x < y$ to represent this order. As explained in Section 1.8, the *precursor symbol* $<$ is analogous to the familiar “less than” symbol $<$ in integer arithmetic, although $<$ is more specific, since $x < y$ means that x *directly* precedes y . In discrete classical causal theory, the local “arrow of time” is defined by this primitive order, with a single fundamental

unit of local time, i.e., a single “classical chronon,” separating x and y . For a larger family D of elements, say $D = \{w, x, y, z\}$, one must consider a corresponding family of relations, encoding each individual instance of cause and effect. It is standard to denote this entire family, say $\{w < y, x < y, z < w\}$, by the single symbol $<$. In technical terms, this means that $<$ is a **binary relation** on D , i.e., a subset of the Cartesian product $D \times D$. Using this definition, the statement $x < y$ means that the ordered pair (x, y) in $D \times D$ is an element of $<$. In this book, the pair $(D, <)$ is called a *directed set*. Equivalent structures are called *directed graphs* in conventional mathematical settings, with the term “directed set” usually assigned a more specific meaning; however, in this book, a “directed set” means simply a set equipped with a binary relation. An individual relation between a pair of elements x and y in D is almost always represented by the notation $x < y$, rather than the alternative notation $(x, y) \in <$.

The “up the page” convention for inferring the directions of relations may be used for any directed set D in which “causal influence always flows one way, never looping back.” Technically, this means that D is *acyclic*, as noted in Section 1.4. All of the classical histories used for illustrative purposes in Chapter 1 are modeled via acyclic directed sets, but it is sometimes interesting to adopt a broader viewpoint, and allow directed sets containing cycles. In spite of their counterintuitive properties, such sets can be physically interesting, partly because general relativity does not rule out the existence of “closed causal curves.” Generalized Hasse diagrams cannot be used to represent such sets; rather, arrows must be included in the diagrams to explicitly indicate the direction of each relation. The vast majority of directed sets considered in this book, however, are acyclic. In Figure 2.2.2, I reproduce the generalized Hasse diagram of the “slightly more complicated classical history” of Figure 1.4.3, now viewed abstractly as an acyclic directed set.

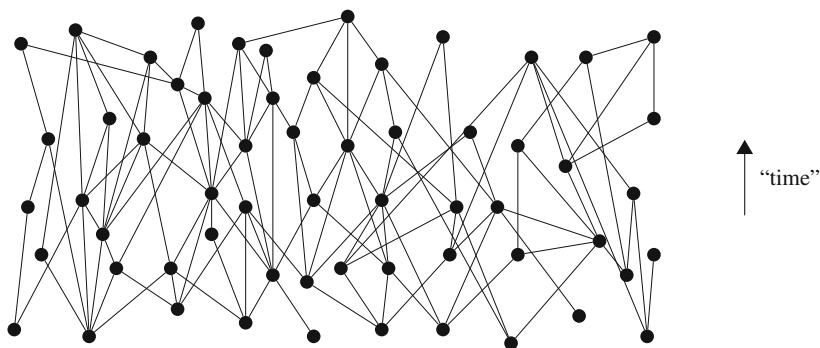


Figure 2.2.2. An acyclic directed set $(D, <)$.

Classical causal metric hypothesis in terms of directed sets. In the classical context, a general mathematical version of the causal metric hypothesis may be formulated in terms of directed sets, as follows:

Definition 2.2.3. Classical causal metric hypothesis (CCMH). *Classical histories may be modeled via directed sets.*

Though more specific than Definition 2.1.1, this is still a very general statement. For example, it includes Sorkin’s version of the causal metric hypothesis, since causal sets are a special case of directed sets, but it also includes continuous *domain-theoretic* versions, since domains are directed sets as well. As noted in Section 2.1, practical application of discrete causal theory requires narrowing the focus to a *specific* class of “physically relevant directed sets,” and this involves nontrivial choices. In this book, *the most substantial such choices are to work in the discrete context, to focus mainly on the acyclic case, and to work primarily in terms of direct, i.e., immediate, relationships.* In technical terms, discreteness is expressed via an appropriate *local finiteness condition*; in this case, *star finiteness* (SF). *Acyclicity* (AC) simply means that no sequence of relations may both begin and end at the same element. The choice to focus on direct relationships is implemented by modeling classical causal structure in terms of generally *nontransitive* binary relations, and interpreting each individual relation to encode independent influence. This means abstaining from the axiom of *transitivity* (TR), and adopting the *independence convention* (IC). These choices are briefly explained below, and are examined and justified more systematically in Chapters 3 and 4.

Discreteness. The single most significant restriction on the types of directed sets studied in this book as models of classical spacetime is that they are *discrete*. The term “discrete” has different meanings in topology, measure theory, and order theory, so further explanation is necessary to render the precise meaning of this choice sufficiently clear. In causal set theory, the axiom of *interval finiteness* (IF) ensures that causal sets are discrete in an order-theoretic sense; i.e., every nonextremal element in a causal set has at least one *maximal predecessor* and *minimal successor*. In addition, the use of a *discrete measure* that “counts fundamental volume units” ensures that causal sets are discrete in a measure-theoretic sense.³ This lends plausibility to the appearance of “number” in Sorkin’s version of the classical causal metric hypothesis, without any need to “quantize spacetime,” as is typically attempted in continuum-based approaches to quantum gravity. The causal set version of discreteness *may* take too literally the idea of “volume” at the fundamental scale, since “familiar geometric notions” are generally expected to emerge only at relatively large scales in discrete causal theory.⁴ For the purposes of this book, the order-theoretic meaning of discreteness, suitably generalized, is the most important. The measure-theoretic meaning, meanwhile, is “relevant, but less precise,” while the topological meaning is almost completely irrelevant. The desired order-theoretic and measure-theoretic

³To be precise, the causal set measure allows “statistical fluctuations” in the assignment of volume. This topic is revisited in Sections 3.2, 3.5, and 4.5.

⁴One such familiar notion that turns out to be very complicated is the emergent notion of spacetime *dimension*. In particular, a variety of approaches to discrete spacetime structure suggest possible *scale-dependence* of dimension. See the recent paper of Carlip [CA15] for an interesting discussion and list of references.

properties may be achieved by imposing the local finiteness condition of *star finiteness* (SF), together with the *generalized measure axiom* (M*), both discussed in Chapter 4. Some of the general motivations for working in the discrete context are discussed further in Sections 2.9 and 2.10, as well as in later chapters.

Acyclicity. Another attractive “physical relevance criterion” for directed sets is acyclicity (AC), which justifies the “up the page” convention for the generalized Hasse diagrams appearing in this book.⁵ Acyclicity is a natural abstraction of the apparently *unidirectional* nature of causality, i.e., the fact that events do not seem to contribute to their own causes, either directly or indirectly. The acyclic binary relations studied in this book are generally *not* assumed to be transitive (TR), since transitive relations cannot distinguish naturally between direct and indirect relationships. By abstaining from transitivity, one is free to use the independence convention (IC), which specifies that each individual relation encodes direct influence, independent of other modes of influence between its initial and terminal elements. A binary relation $<$ on a set D generates a *transitive relation* $<_{\text{tr}}$ under the operation of *transitive closure*, as described in Chapter 3. In the acyclic case, $<_{\text{tr}}$ is a *partial order*, which is essentially why partial orders play such a prominent role in causal set theory. In particular, this lends plausibility to the appearance of “order” in Sorkin’s version of the classical causal metric hypothesis. However, as indicated above, passage from $<$ to $<_{\text{tr}}$ destroys information about direct relationships; i.e., one cannot recover $<$ from $<_{\text{tr}}$. For this reason, I work almost exclusively with the nontransitive binary relation $<$ itself, which I call the *causal relation*. These details are discussed in Chapter 3, particularly in Sections 3.9 and 3.10.

2.3 Structure on Relativistic Spacetime

“Geometric intuition” versus physical geometry. Given the “geometric” picture of directed sets afforded by graphical representations such as generalized Hasse diagrams, it is natural to view these sets as “spaces” in a *mathematical* sense, quite apart from any specific physical interpretation. Of course, I have already explained the discrete causal interpretation of these sets at an informal level, but this preliminary description falls far short of providing a precise quantitative description of how discrete causal theory models fundamental spacetime structure, or enabling a meaningful comparison of the theory with more conventional approaches to fundamental physics. This situation calls for a deeper examination of the physical role of directed structure, followed by a careful explanation of its consequences in the discrete setting. The best understanding of such structure available within the scope of experimentally-established physics comes from general relativity. Hence, much of the present chapter focuses on reviewing a few important aspects of relativistic

⁵For diagrams of directed sets that are *not* assumed to be acyclic, arrows are added to the edges to explicitly indicate the direction of causal influence. Examples appear in Figures 2.7.2, 3.6.5, 5.4.4, and 8.7.2. It is occasionally convenient to add arrows even in the acyclic case.

spacetime structure, from a perspective that may be readily adapted to the discrete causal context. Central to this picture are the *metric recovery theorems*.

The pedestrian view of directed sets as “spaces in a mathematical sense” represents an instance of the common and useful practice of applying “geometric intuition” to help analyze the properties of mathematical objects, whether or not these objects are “geometric” in a traditional sense. While the original source of such intuition is often partly physical in nature, the practice itself has no necessary connection to physics at all; indeed, fields of pure mathematics such as functional analysis and algebraic geometry abound with such “geometric” methods. In functional analysis, for example, one studies “spaces of functions,” which are typically infinite-dimensional vector spaces. The elements of such spaces are functions on some other space, such as the real line \mathbb{R} . Notions originating in geometry, such as projections and orthogonality, play a central role. In algebraic geometry, meanwhile, one studies “spaces” called *algebraic schemes*, whose elements are prime ideals in commutative rings; for example, the ideal of all polynomials $f(x, y)$ in the polynomial ring $\mathbb{R}[x, y]$ vanishing on an irreducible algebraic curve⁶ such as $\{(x, y) \in \mathbb{R}^2 | y = x^2\}$. In this case, familiar “geometric” concepts are applied in ways unimagined by mathematicians over the first two millennia of studying such objects; for example, individual points may possess nonzero dimension. Generalization of these ideas to the noncommutative setting leads back in a curious manner to topics in fundamental physics, via Connes’ *noncommutative geometry*, revisited briefly in Chapter 8.

The purpose of rehearsing this bit of pure mathematics is to emphasize that the vague “geometric” character of arbitrary directed sets does not, by itself, constitute evidence that physical spacetime, from which humans have acquired much of their geometric intuition, may actually *be* a directed set, or a structure “built from directed sets.” Indeed, “geometric data” in some form may be squeezed out of almost any type of mathematical object one might choose to work with. For example, when studying any suitable class⁷ of structured sets, one may always pass to a *category* of such sets, then ignore the “internal structure” of the sets themselves, regarding them as merely “higher-level elements,” just as directed sets are viewed as “elements” of a kinematic scheme. This yields an abstract *multidirected set*, whose elements represent the original structured sets, and whose relations represent morphisms between pairs of structured sets. This multidirected structure, in turn, provides natural “geometric” notions of “directions,” “neighbors,” “paths,” “distances,” and so on. Of course, neither multidirected sets nor categories enter the picture in any serious manner until Chapter 3, aside from a brief explanation in the present section regarding the role of category theory in organizing different types of structure on relativistic spacetime. However, the details of this particular example involving categories of structured sets are immaterial at present. Its role is merely to illustrate why the causal metric

⁶In this context, the word “curve” means “locus of points,” *not* “map from a real interval into a manifold,” as it does later in the chapter.

⁷The reason for the qualifier “suitable” here is that the class must be “small enough” so that the resulting multidirected structure will actually be a set, rather than a *proper class*.

hypothesis (CMH) requires much stronger and more specific justification than I have demonstrated thus far.

Fortunately, such justification exists, in the form of the metric recovery theorems of Hawking and Malament. The basic idea of metric recovery is that *almost all of the apparent geometric structure of relativistic spacetime is encoded in its causal structure*. The only information missing is “scale data;” or, more precisely, a *conformal factor* in the metric. If spacetime is actually discrete, however, then the combinatorial details of discrete microstructure can supply scale data “for free.” Hence, one may construct discrete models whose *only* structure is causal structure, yet which “look just like relativistic spacetime at ordinary scales.” This suggests that, under the limitations of present observations, discrete causal models of classical spacetime are “just as good” as the geometric models used in relativity. In fact, they turn out to be much better in a number of significant ways. The principal reason why such models are not yet ready to replace relativistic spacetime root and branch is because general relativity explains *how* specific geometry arises dynamically, while discrete causal dynamics is still in its infancy.

Pseudo-Riemannian manifolds; diffeomorphism invariance. To properly understand the subject of metric recovery, it is necessary to examine a few of the geometric ingredients of general relativity. In this context, the “spaces” of interest are special types of *real manifolds*, called *pseudo-Riemannian manifolds*, viewed as models of classical spacetime. To be precise, I should point out that even in general relativity, such manifolds are not properly regarded as “physical” in their own right. Einstein himself understood that individual elements of a pseudo-Riemannian manifold do not possess intrinsic physical meaning, as already mentioned in Section 1.6. To understand, at an informal level, why this is true, one might imagine “painting” certain physical information on the surface of a sphere, then mapping each element of the sphere to another element by means of a rotation. In this context, the physical information is “re-associated” with different abstract points on the sphere, but there is no intrinsic physical distinction between the two associations.⁸ More formally, general relativity is *diffeomorphism invariant*; i.e., Einstein’s equation (1.3.1) does not change its form under “smooth transformations” of relativistic spacetime. With this understanding, I often lapse into the common habit of treating elements of a pseudo-Riemannian manifold as “spacetime events” in the relativistic context, even though these elements really only *represent* spacetime events.

This distinction between mathematical elements and physical events actually turns out to be important when comparing general relativity to discrete causal theory, due to the relative *rigidity* of discrete directed sets, mentioned periodically throughout Part I of the book, and revisited more thoroughly in Section 6.3. I will not go into the details in the present chapter, but one of the basic physical implications of this rigidity is that physical spacetime events may be associated *much more directly* with elements of a discrete directed set than with elements of a pseudo-Riemannian manifold. On a historical note, the “non-physicality” of elements and coordinate systems presented Einstein with such severe difficulties that it contributed to several

⁸This familiar thought experiment paraphrases part of an analogous discussion in Rovelli [RO04].

years of delay in the publishing of his first papers on general relativity, even after most of the mathematical and physical essentials were in place. Had Einstein been working with discrete directed sets instead, this particular conceptual issue would likely have posed far less of an obstacle.⁹

Five types of structure. Pseudo-Riemannian manifolds are endowed with a number of different types of structure, with varying types and degrees of physical significance. Two manifolds that are “the same” with respect to one type of structure may be “different” with respect to another. Five important types of structure on a pseudo-Riemannian manifold X are *topological*, *smooth*, *causal*, *conformal*, and *metric* structure. These are listed in a suggestive way in Figure 2.3.1. For future reference, the left-hand side of the figure, which looks like part of a curved two-dimensional surface, really represents part of a pseudo-Riemannian manifold, usually assumed to be four-dimensional, connected, and without boundary. In particular, the “edges of the surface” do not represent actual boundary points of the manifold, but merely delimit the portion being illustrated. Auxiliary structure represented by graphical features that intersect the edges, such as “curves drawn on the surface,” should be assumed to “keep on going,” rather than “stopping at the edge.”

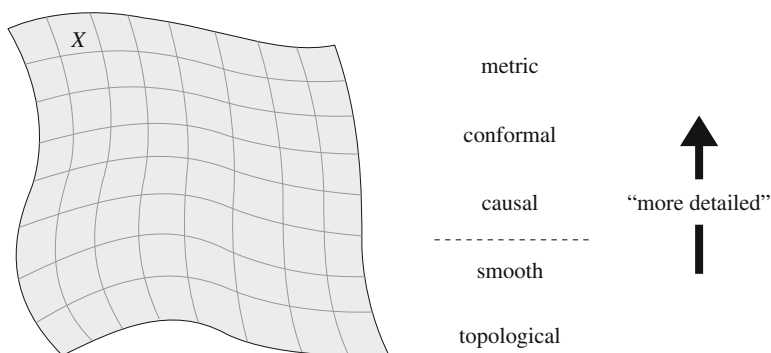


Figure 2.3.1. Informal view of types of structure on a relativistic spacetime manifold X .

“Relativistic spacetime geometry” is an informal term for metric structure in relativity, which is the “most detailed” of the five types of structure listed in the figure. A question of central importance, both in relativity and in related theories, is how much of this metric structure one may “recover” from knowledge of one or more of the other four types of structure. This is a generalized version of the metric recovery problem. In the present context, of course, the focus is directed toward the more specific question considered by Malament; namely, how much of the *metric* structure may be recovered from knowledge of the corresponding *causal* structure. However, these five types of structure are intimately bound together, and it

⁹Rovelli [R004] gives a splendid explanation of this topic in his chapter 2.

is necessary to consider all of them to some degree in order to understand this more specific problem.

Ranking of structures. The list appearing on the right-hand side of Figure 2.3.1 gives an informal “ranking” of these five types of structure in the specific context of relativistic spacetime. The qualitative idea of this ranking is that it is “easy” to recover structures lower on the list from structures higher on the list, but difficult or impossible to do the opposite. For example, metric structure is specified by a particular choice of metric, as described below, while conformal structure is specified by an equivalence class of metrics, related to each other by “scaling functions” called *conformal factors*. Knowledge of metric structure, i.e., of a specific metric, immediately yields knowledge of the corresponding equivalence class of metrics, i.e., of the conformal structure. However, knowledge of an equivalence class of metrics, by itself, does not yield a method of choosing a unique metric from among them. Hence, it is generally impossible to recover metric structure from conformal structure alone.

The reason for the dashed line in the figure is that it is possible to talk about the top three types of structure in “non-smooth” contexts; in particular, *discrete* causal structure is the main subject of Part I of this book. Hence, the “ranking” only applies to relativistic spacetime, where an underlying smooth manifold structure is taken for granted. Such a manifold comes equipped with a “standard” topological structure, called the *manifold topology*, indirectly inherited from the topological structure of the real line \mathbb{R} . However, other topological structures are also of interest in relativity; for example, the *Alexandrov topology*, which is closely related to the axiom of interval finiteness (IF) in causal set theory, and the *path topology* of Hawking, King, and McCarthy [HA76]. The latter two topologies are “physically more natural,” in certain specific ways, than the manifold topology, and both contribute to the proof of the metric recovery theorems. However, when I refer to topological structure on relativistic spacetime in this chapter, I am referring to the manifold topology unless stated otherwise.

For the sake of relevance and brevity, I focus primarily on the top three types of structure listed in the figure, examining them in reverse order of detail, beginning with metric structure and ending with causal structure. The choice to present the material in this order is motivated by the fact that conformal structure and causal structure in relativity are usually expressed *in terms of* metric structure. I discuss smooth structure and topological structure only briefly, and in purely auxiliary ways. For example, smooth structure enters the picture in the discussion of metric structure, since the bilinear maps on tangent spaces defined by the metric are taken to vary smoothly. It also plays a role in the discussion of causal structure, since the distinction between *causal isomorphisms* and *enhanced causal isomorphisms* in Definition 2.6.3 is described in terms of the distinction between smooth causal curves and causal curves that are merely continuous. I discuss topological structure last of all, and only briefly, after the description of metric recovery in Section 2.8. The purpose of this brief topological detour is partly to aid the reader in understanding the literature on metric recovery, which makes use of all three topologies mentioned above, and

partly to facilitate the discussion in Chapter 4 regarding how the Alexandrov topology relates to the interval topology and the causal set axiom of interval finiteness (IF).

The qualitative claim of the classical causal metric hypothesis (CCMH) is that “metric structure” is merely *a way of describing certain aspects of causal structure*. This assertion seems to turn on its head the ranking of structures appearing in Figure 2.3.1, but the causal metric hypothesis is not intended to apply, at a precise level, in the relativistic setting. Indeed, if classical spacetime *really is* precisely represented by a pseudo-Riemannian manifold in the usual relativistic sense, then the classical causal metric hypothesis is *wrong*; at least, under the standard definitions and assumptions specifying how causal structure should be modeled in this context. In this case, the metric recovery theorems tell exactly what is missing; namely, scale data, and this is where conformal structure enters the picture. Hence, the plausibility of the causal metric hypothesis depends on the conviction that what is conventionally viewed as relativistic spacetime is *really some alternative structure that merely mimics a pseudo-Riemannian manifold at large scales*. Although the metric recovery theorems say nothing about discreteness explicitly, they do demonstrate indirectly that “relativistic spacetime looks suspiciously like a discrete directed set,” since discrete directed structure carries natural scale data. These theorems may therefore be interpreted to suggest that discrete directed sets represent a particularly promising candidate for such alternative structure. It is therefore reasonable to consider the possibility that general relativity is merely a “smooth approximation” of discrete causal theory, just as special relativity is a “flat approximation” of general relativity, and Newtonian mechanics is a “low-velocity approximation” of special relativity.

Category theory as an organizing principle. Category theory provides a useful, though incomplete, method of organizing the five types of structure listed in Figure 2.3.1. It is useful because it treats these structures in a unified and coherent manner, but it is incomplete because it emphasizes only the *active viewpoint*, in which structural relationships are studied via *morphisms*, in this case, by actively mapping each element of spacetime to another element. The complementary *passive viewpoint*, in which one compares different instances of the same type of structure without any active mapping procedure, is implicitly deprecated in this context. In conventional continuum-based physics, these two viewpoints often, though not always, involve essentially equivalent treatment of physically relevant information. In discrete causal theory, however, they lead in profoundly different directions. I return to this subject later in the chapter, and again in Section 8.4. At present, the distinction introduces no serious difficulties.

If X and X' are “spaces,” in a general and unspecified sense, with each space possessing different instances of a particular type of structure, then X and X' may be viewed as objects in a category characterized by this common structural type, ignoring for the moment any other type of structure they might possess. For example, if X and X' are “relativistic spacetimes,” then one may choose to view them as simply topological spaces, i.e., as objects in an appropriate *topological category*. For the sake of clarity, it is useful to note that there is a different category for each type of topology; for example, the *manifold*, *Alexandrov*, and *path topologies*.

Alternatively, one may choose to view X and X' as smooth real manifolds, i.e., as objects in an appropriate *smooth category*. Smooth real manifolds are the underlying structures on which geometry is built in general relativity. Once a particular type of structure has been chosen to study, it is natural to turn attention to the class of maps $f : X \rightarrow X'$ preserving this structure. These are the *morphisms* between X and X' in the chosen category. For more abstract categories, morphisms are not necessarily maps, but in the context of relativity, the categories involved are very “concrete,” and maps suffice. If the morphism $f : X \rightarrow X'$ possesses an inverse that is also a morphism, then f is called an *isomorphism*, and X and X' are called *isomorphic*. This is sometimes denoted more succinctly by the expression $X \cong X'$, which expresses the information that there exists *at least one* isomorphism between X and X' , without specifying a *particular* isomorphism. Isomorphic objects X and X' are considered to be “essentially the same” with respect to whatever type of structure is being studied. “Self-morphisms” $f : X \rightarrow X$ are called *endomorphisms*, and “self-isomorphisms” are called *automorphisms*. It is easy to see why this approach emphasizes only the active viewpoint, since different instances of a particular type of structure on X are compared by *actively transforming* X .

Traditionally, different names are assigned to morphisms in different categories. For example, a topological isomorphism is called a *homeomorphism*, and one must specify which type of topological structure is being considered for this notion to be well-defined. A smooth isomorphism is called a *diffeomorphism*, a metric isomorphism is called an *isometry*, and a conformal isomorphism is called a *conformal isometry*. As far as I know, causal morphisms do not possess separate traditional names, probably because they were not seriously studied until after category theory became the standard structural paradigm in abstract algebra. In particular, Zeeman [ZE64] seems to have coined the term *causal automorphism* in his 1964 paper on causality and the Lorentz group.¹⁰ In the topological category for the manifold topology, relativistic spacetime is assumed to be locally homeomorphic to \mathbb{R}^4 ; i.e., it is a four-dimensional real manifold. In the more detailed smooth category, it is assumed to be locally diffeomorphic to \mathbb{R}^4 with its usual smoothness structure; i.e., it is a *smooth* four-dimensional real manifold. This supplies enough underlying structure to facilitate the specification of spacetime geometry.

2.4 Metric Structure

Pseudo-Riemannian metrics. The type of geometry of principal interest in general relativity is a special kind of *pseudo-Riemannian geometry*, sometimes called *Lorentzian geometry*. It is the tool Einstein finally settled on for modeling classical spacetime structure after several years of painful self-education in then-relatively-

¹⁰In fact, both Zeeman [ZE64] and Malament [MA77] define causal morphisms in terms of *timelike* rather than *causal* relationships. The latter may be *either* timelike or null. Discrete causal theory generally does not make such distinctions. See Sections 2.6 and 2.8 for more details.

modern mathematics. A significant proportion of readers will probably be grateful for the inclusion of a few extra pages recalling some of the rudiments of this particular type of geometry. However, I do assume that the reader is familiar with the basic definitions of *real manifolds*, *tangent* and *cotangent spaces*, *smooth maps*, and a few related ideas from elementary differential geometry.

Definition 2.4.1. *Let X be a smooth real manifold. A **pseudo-Riemannian metric** g on X is a smoothly-varying family of real-valued, non-degenerate, symmetric, bilinear maps on the tangent spaces of X . A **pseudo-Riemannian manifold** is a smooth real manifold together with a choice of pseudo-Riemannian metric.*

To spell this out in more detail, the definition means that for each $x \in X$, g assigns a real value $g_x(v, w)$ to each pair of tangent vectors v and w in the tangent space $T_x X$ at x . For any fixed tangent vector v in $T_x X$, one may define a map $g_x(v, -) : T_x X \rightarrow \mathbb{R}$, sending each tangent vector w to $g_x(v, w)$. A similar map $g_x(-, w) : T_x X \rightarrow \mathbb{R}$ may be defined by fixing the second argument in g_x . The “non-degenerate” property of g means that the map $g_x(v, -)$ is identically zero if and only if v is itself the zero vector, and similarly for $g_x(-, w)$. The “symmetric” property means, of course, that $g_x(v, w) = g_x(w, v)$ for every choice of v and w . The “bilinear” property means that g_x is linear in each of the variables v and w . For the first variable v , this means that $g_x(a_1 v_1 + a_2 v_2, w) = a_1 g_x(v_1, w) + a_2 g_x(v_2, w)$, for any tangent vectors v_1, v_2 , and w at x , and any scalars a_1 and a_2 . An alternative way to say this is that the maps $g_x(v, -)$ and $g_x(-, w)$ are linear; i.e., they are *dual vectors* or *cotangent vectors* at x .

The metric g is a *tensor*, which is a general term denoting a family of multilinear maps of an appropriate type, whose arguments are tangent vectors and/or cotangent vectors on X . The assignment $(v, w) \mapsto g_x(v, w)$ may be viewed as a “generalized inner product” on the vector space $T_x X$. In the special case of a *Riemannian* manifold, dropping the prefix “pseudo,” this assignment really *is* an inner product, i.e., a symmetric, bilinear, positive-definite map, where “positive” means that $g_x(v, v) \geq 0$, and “definite” means that $g_x(v, v) = 0$ if and only if $v = 0$. More generally, however, it is possible that $g_x(v, v) \leq 0$ even when the vector v is nonzero; in this case, g does not define true inner products on the tangent spaces of X . This occurs, in particular, in the relativistic case, where the sign of $g_x(v, v)$ determines whether v is *timelike*, *null*, or *spacelike*. These designations, along with their physical interpretations, are discussed further below. For notational clarity, I remark that it is sometimes convenient to denote a pseudo-Riemannian manifold by a pair (X, g) , when one wishes to make the choice of metric explicit.

In the context of relativity, one is interested in the specific case of four-dimensional spacetime. In this case, the metric g may be represented at each point $x \in X$ by a 4×4 symmetric matrix:

$$(g_{\mu\nu}) = \begin{pmatrix} g_{00} & g_{01} & g_{02} & g_{03} \\ g_{10} & g_{11} & g_{12} & g_{13} \\ g_{20} & g_{21} & g_{22} & g_{23} \\ g_{30} & g_{31} & g_{32} & g_{33} \end{pmatrix}. \quad (2.4.2)$$

The numerical values of the entries $g_{\mu\nu}$ in this expression depend on a choice of basis for $T_x X$, which is often derived from a choice of local coordinates on X near x . However, the metric g itself, as opposed to a particular matrix representation of g at a particular point, is often written as $g_{\mu\nu}$, for historical reasons. For example, this notation appears in the usual expression of Einstein's equation (1.3.1). Mathematically, this is a bit awkward, and constitutes one of the reasons why much of the literature on general relativity, and also on quantum field theory, is difficult for many mathematicians to read. A reasonable compromise between traditional and modern conventions is Penrose's *abstract index notation*, in which the indices appearing in the expression for a tensor have nothing to do with bases or coordinates, but merely indicate the type and order of its arguments, i.e., the number and arrangement of tangent vectors and cotangent vectors on which it operates. However, such notational details play essentially no role in this book beyond the present chapter.

Examples of metrics. The typical student of relativity usually encounters two specific, and particularly simple, pseudo-Riemannian metrics, before studying the properties of metrics in general. The first of these metrics is the *Euclidean metric* δ on \mathbb{R}^4 , for which the diagonal entries in the above matrix representation (2.4.2) are 1, 1, 1, 1, and the off-diagonal entries are 0, for every point $x \in \mathbb{R}^4$, under a standard choice of basis. The second is the *Minkowski metric* η on \mathbb{R}^4 , for which the corresponding diagonal entries are $-1, 1, 1, 1$, and the off-diagonal entries are 0. Since the generalized inner products defined by these metrics do not vary across spacetime, they are called *constant metrics*. In both cases, the entire pseudo-Riemannian manifold involved is isomorphic to any of its tangent spaces. This means that properties that generally apply only locally, or in a limiting sense, such as *Lorentz invariance* in the case of relativistic spacetime, actually hold at a global level in these special cases. A simple example of a non-constant metric is the *Schwarzschild metric* on \mathbb{R}^4 , which may be represented by the matrix

$$(g_{\mu\nu}) = \begin{pmatrix} -\left(1 - \frac{2GM}{c^2 r}\right) & 0 & 0 & 0 \\ 0 & \left(1 - \frac{2GM}{c^2 r}\right)^{-1} & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & r^2 \sin^2 \theta \end{pmatrix},$$

using coordinates (ct, r, θ, ϕ) , where c is the speed of light, G is Newton's gravitational constant, t is the time coordinate, and the spatial coordinates (r, θ, ϕ) are the usual spherical coordinates on \mathbb{R}^3 .¹¹ In the limit of a vanishing “cosmological constant,” the Schwarzschild metric describes relativistic spacetime near an appropriate spherically-symmetric body of mass M , such as an ideal, non-charged, non-rotating black hole.

¹¹ It is common to “choose units” in such a way that the numerical values of constants such as c and G are set to 1. For example, Carroll [CA04], p. 193, omits explicit inclusion of c in his expression of the Schwarzschild metric, and Hawking and Ellis [HE73], p. 149, omit both c and G . I include these factors so that the “units work out” in a naïve sense.

A general class of metrics of particular interest in conventional relativistic cosmology is the class of *Friedman-Lemaître-Robertson-Walker* (FLRW) metrics, which are special solutions to Einstein’s equation (1.3.1), describing homogeneous, isotropic spacetimes with time-varying “scale factors.” Under a suitable choice of coordinates (ct, r, θ, ϕ) , a FLRW metric may be represented by the matrix

$$(g_{\mu\nu}) = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & a^2(t) & 0 & 0 \\ 0 & 0 & a^2(t)f^2(r) & 0 \\ 0 & 0 & 0 & a^2(t)f^2(r)\sin^2\theta \end{pmatrix},$$

where $a(t)$ is the scale factor, and where $f(r)$ takes on one of a few simple values, depending on the curvature. Two specific “maximally symmetric” examples of particular prominence are the positive-curvature *de Sitter metric*, which describes an idealized cosmology dominated by a positive “cosmological constant,” and the negative-curvature *anti-de Sitter metric*, whose higher-dimensional analogues are central to Maldacena’s *AdS/CFT correspondence* [MA99] in string theory. The most popular “realistic” models in modern cosmology involve so-called “almost FLRW metrics,” which are perturbed away from an exact FLRW form by inhomogeneities on relatively small scales, in an effort to model the manifest local irregularity of matter content in the observable universe. In particular, the Λ CDM model, named to reflect the fact that it involves a nonzero “cosmological constant” Λ and “cold dark matter” (CDM), uses such metrics. A good standard reference on the subject is [EMM12].¹² FLRW models make a brief reappearance in Chapter 4, where I quote an assertion by the founders of causal set theory [BLMS88] that star finite causal sets suffice for studying their discrete causal analogues.

Pulling back a metric; metric morphisms. Given a smooth morphism¹³ $f : X \rightarrow X'$, between a pair of smooth real manifolds X and X' , together with a choice of pseudo-Riemannian metric g' on X' , one may define a smoothly-varying family of symmetric, bilinear maps f^*g' on the tangent spaces of X , called the *pullback* of g' along f . This family is defined by “pushing forward” tangent vectors from X to X' , then applying g' to these tangent vectors. If the morphism f possesses “suitable properties,” in particular, if it is a diffeomorphism, or more generally, an *immersion*, then f^*g' is nondegenerate, and therefore constitutes a pseudo-Riemannian metric on X . The reason for such a brief and informal description of this construction is that the pullback plays a very limited and specific role in this chapter. Most of the technical details are unnecessary for this purpose, but may be found in any appropriate reference on differential geometry or general relativity, such as Wald [WA84]. A *metric morphism* between a pair of pseudo-Riemannian manifolds (X, g) and (X', g') is a map $f : X \rightarrow X'$ that preserves metric structure, in the sense that the pulled-back metric f^*g' on X coincides with the chosen metric g on X . In particular, if $X = X'$,

¹²See [EMM12], p. 205. Most texts based on general relativity use the “line element” notation to describe such metrics; in this case, $ds^2 = -dt^2 + a^2(t)[dr^2 + f^2(r)(d\theta^2 + \sin^2\theta d\phi^2)]$.

¹³This means a morphism in the smooth category.

then such a morphism f is necessarily bijective, and is therefore a self-isometry, i.e., an automorphism of X in the category of pseudo-Riemannian manifolds.

Metric signature. The *signature* of a pseudo-Riemannian metric g on a smooth n -dimensional real manifold X is an ordered pair of natural numbers (p, q) , with sum n , where p and q are the “numbers of plus 1’s and minus 1’s along the diagonal,” when g is represented in matrix form, via appropriate choices of bases for the tangent spaces $T_x X$. Again, an informal description suffices for the present purposes. The signature of the Euclidean metric δ on \mathbb{R}^4 is $(4, 0)$, since the corresponding diagonal entries are 1, 1, 1, 1 in this case, while the signature of the Minkowski metric η on \mathbb{R}^4 is $(3, 1)$, since the diagonal entries are $-1, 1, 1, 1$.¹⁴ More generally, signatures of the form $(p, 1)$, i.e., signatures with “exactly one minus sign,” are so important in theoretical physics, regardless of the dimension of the underlying manifold, that they are collectively referred to by the single name **Lorentz signature**. Other metric signatures may also be physically relevant; for example, the signature $(2, 2)$ plays a role in Penrose’s *twistor theory*. The fact that the signature of g is independent of the choice of point $x \in X$ and the choice of basis for $T_x X$, is due to *Sylvester’s law of inertia*, which says that “the number of positive and negative coefficients in a diagonalized quadratic form does not depend on the choice of diagonalization,” together with the “smoothly varying” hypothesis on g .

Riemannian geometry is the study of real manifolds of metric signature $(p, 0)$, for some p , called *pure signatures*. In this sense, Riemannian geometry is a generalization of Euclidean geometry, since the spaces involved “look like Euclidean spaces locally,” in a limiting sense, even though their underlying smooth manifold structures may be very complicated. **Pseudo-Riemannian geometry** expands this picture further, to allow *mixed signatures*, i.e., signatures for which both p and q are nonzero; the case $q = 1$ is the Lorentzian case discussed above. Altogether, there exist $n + 1$ possible signatures (p, q) for a pseudo-Riemannian metric on a smooth real manifold of dimension n , running from $(n, 0)$ to $(0, n)$. In particular, in the special case where the manifold under consideration is merely \mathbb{R}^n , there exist $n + 1$ different *pseudo-Euclidean spaces* \mathbb{R}^{p+q} , whose metrics are the constant metrics of signatures (p, q) . This is why Minkowski spacetime is often denoted by \mathbb{R}^{3+1} . Physically, this choice of notation emphasizes the fact that one of the dimensions, viewed here as the “temporal dimension,” is distinguished from the others, due to its association with the single minus sign in the metric signature. More generally, it is easy to understand why Lorentz signature is so important from the perspective of causal structure: the single minus sign in the signature corresponds to the “local direction from cause to effect,” regardless of the dimension of the underlying manifold. For a more general pseudo-Riemannian manifold X with signature (p, q) , the tangent spaces $T_x X$ at each point $x \in X$ are naturally isomorphic to \mathbb{R}^{p+q} as pseudo-Euclidean spaces, so these manifolds are “locally pseudo-Euclidean of type (p, q) ,” in a limiting sense. In

¹⁴The inconsistency of the order of plus and minus signs between the list of entries $-1, 1, 1, 1$ and the abbreviation $(3, 1)$ for the metric signature is an annoying historical artifact. One *ought* to either flip the signs to $1, -1, -1, -1$ and denote the signature by $(1, 3)$, as is done in certain texts on quantum field theory, or else flip the entries to $1, 1, 1, -1$, and denote the signature by $(3, 1)$.

particular, the tangent spaces of pseudo-Riemannian manifolds of Lorentz signature $(3, 1)$ are isomorphic to \mathbb{R}^{3+1} , so relativistic spacetime is “locally approximated by Minkowski spacetime.”

Recovering “physically relevant metric structure.” Among all pseudo-Riemannian manifolds of Lorentz signature $(3, 1)$, those actually arising in “physically relevant scenarios” in general relativity are sometimes given special names. For example, *vacuum* solutions to Einstein’s equation, in which the stress-energy tensor vanishes, are called *Einstein manifolds*. Minkowski spacetime is the prototypical example. The class of Einstein manifolds is too restrictive for the study of metric recovery in the relativistic setting, since the matter-energy content of the observable universe is non-negligible. On the other hand, certain classes of solutions to Einstein’s equation involve types of “exotic matter” whose existence is doubtful, or configurations of matter and energy that may be difficult or impossible to achieve dynamically, even if they are theoretically possible. Hence, it can sometimes be desirable to restrict attention to the recovery of a smaller class of manifolds than the class of *all* solutions to Einstein’s equation for *all* possible configurations of matter and energy. It is convenient to refer to such “physically relevant” manifolds as *generalized Einstein manifolds*. The following definition is deliberately vague, simply because the actual results necessary to motivate the developments in this book are much more general.

Definition 2.4.3. *A pseudo-Riemannian manifold of Lorentz signature $(3, 1)$, satisfying Einstein’s equation (1.3.1) for a “physically reasonable” choice of stress-energy tensor, is called a **generalized Einstein manifold**.*

General relativity is very successful experimentally, and any theoretical effort to improve upon it must eventually reproduce its empirical success. This is a very demanding task, involving detailed quantitative behavior across scales from the everyday scale up to at least the stellar scale, and quite possibly to the scales of “dark matter,” “dark energy,”¹⁵ and beyond. For this reason, the most promising new theories of fundamental spacetime structure are those that naturally approximate the content of general relativity in a comprehensive manner, rather than attempting to reproduce a host of experimental results in a purely coincidental way. This is one reason why “theories” such as *modified Newtonian dynamics* (MOND) are problematic; such approaches may explain a limited range of phenomena quite well, but the scope of explanation of general relativity is so great that it is difficult to imagine a successful replacement for the theory that is not intimately connected to it at a deep structural level. Despite the crucial common thread of causal structure, discrete directed sets might a priori be expected to prove absurdly inadequate for this purpose, since smoothness is one of the basic properties on which pseudo-Riemannian geometry is built. However, there is “plenty of room” for such models to converge with relativity somewhere between the hypothesized fundamental scale and the scales

¹⁵As in the case of the “cosmological constant” and “dark matter,” the quotation marks serve to warn the reader that the term “dark energy” itself suggests a conventional interpretation of certain observed phenomena.

accessible to present-day experiment. The crucial requirement, then, is not that the fundamental constituents of a new theory must match those of relativity in every respect, but that the new theory must adequately approximate relativistic spacetime structure in a uniform and natural way across a suitable range of scales. These considerations narrow the general problem of describing physical spacetime in terms of causal structure to the following much more specific problem:

Relativistic metric recovery problem: *Can directed sets, and preferably discrete directed sets, adequately approximate a suitable class of generalized Einstein manifolds at sufficiently large scales?*

If the answer to this metric recovery question were *negative*, then the classical causal metric hypothesis (CCMH) would be in serious jeopardy. Indeed, general relativity would have to be basically wrong across a broad range of observable scales for the hypothesis to be true. This is not out of the realm of possibility; for example, many physicists have questioned whether “dark matter” and/or “dark energy” might actually represent MOND-like dynamical deviations from Einstein’s equation. At present, however, it is unnecessary to explore these issues further, because the metric recovery theorems solve a much more general problem:

Solution: *Discrete directed sets can adequately approximate the entire class of four-dimensional¹⁶ pseudo-Riemannian manifolds of Lorentz signature at sufficiently large scales.*

The next four sections of the chapter explain the meaning and significance of this solution.

2.5 Conformal Structure

Spacetime “angles” and “scales.” Before stating an appropriate version of the metric recovery theorems, I must supply some preliminary information about conformal structure and causal structure in general relativity. To avoid the nuisance of copying lists of technical properties, and becoming bogged down in mostly irrelevant discussions about which specific properties apply in which cases, I will sometimes refer to the manifolds involved as merely “relativistic spacetime manifolds,” or even just “spacetimes,” in what follows, even though they may not be actual solutions to Einstein’s equation. A more descriptive term might be “relativistic classical histories,” but this term is not quite accurate, due to the imperfect background independence of general relativity, i.e., the fact that typical relativistic scenarios include *material content* distinct from pure geometry.

¹⁶The same is true, in fact, for any dimension at least three, as discussed in Section 2.8.

The metric g on a relativistic spacetime manifold X encodes information that enables measurement of several different types of geometric quantities in X , such as “angles” and “scales.” The reason for the quotation marks here is that a mixed signature, in particular, the Lorentz signature $(3, 1)$, alters the naïve Euclidean picture of such quantities. For example, the angle between two smooth intersecting curves in a Euclidean space, and by extension, in a Riemannian manifold X , is defined, reasonably enough, to be the angle between their tangent vectors at their point x of intersection. This angle, in turn, is defined in terms of the inner product on the tangent space $T_x X$, which is supplied by the metric. However, if X is a pseudo-Riemannian manifold of mixed signature, then the metric g does not define true inner products on the tangent spaces of X , and this makes the picture subtler. In particular, this is what leads to consideration of *analogues* of Euclidean angles in relativity, such as the *hyperbolic angles* measuring the “rapidity of reference frames in relative motion.” An iconic feature of mixed signature, which distinguishes it from the Riemannian case, is the existence of *null vectors* in the tangent spaces $T_x X$, i.e., nonzero vectors v that are “orthogonal to themselves,” in the sense that $g_x(v, v) = 0$. In the special case of Minkowski spacetime \mathbb{R}^{3+1} , these vectors represent the trajectories of light rays, and define the “light cone,” or **null cone**, of x in \mathbb{R}^{3+1} .¹⁷ Despite such distinctions, the intuition associated with Euclidean “angles” and “scales” remains valuable, and I make informal use of these concepts in some of the examples and illustrations below. The purpose of these examples is merely to *motivate* the notions of “conformal equivalence of metrics” and “conformal maps between spacetimes.” Hence, little precision is needed.

Separating “angle data” and “scale data.” The information encoded in the metric g on a relativistic spacetime manifold X may be partitioned in various ways; in particular, one may study “angle data” and “scale data” separately. To understand how these two types of information may be distinguished, it is instructive to consider a pair of smooth curves γ_1 and γ_2 in X that “intersect with angle θ ” at a point $x \in X$, as illustrated in the left-hand diagram in Figure 2.5.1. There are several choices for how to make this scenario more precise, if one wishes to do so, and any of these choices serve adequately for the purposes of illustration. For example, one may choose to take X to be a Riemannian manifold in this particular example, and view the “angle” between the curves at x as an actual angle; or one may ignore the “up the page rule,” and think of the figure as representing a “spacelike section” of X , which possesses a natural “induced Riemannian structure.” Finally, one may take the tangent vectors of the curves to be timelike, in the sense described below, and view this “angle” as a hyperbolic angle.

To illustrate the distinction between “angle data” and “scale data” on X , one may replace the metric g with a new metric $g' = \Omega^2 g$, for some positive real

¹⁷ As explained in Section 4.5, the “near-zero” Minkowski spacetime intervals between an event and other events near its null cone in a given frame of reference translates to extreme *spatiotemporal nonlocality* in certain idealized types of causal sets induced by global sprinklings into \mathbb{R}^{3+1} . This leads to interesting general considerations regarding *local structure* in discrete causal theory.

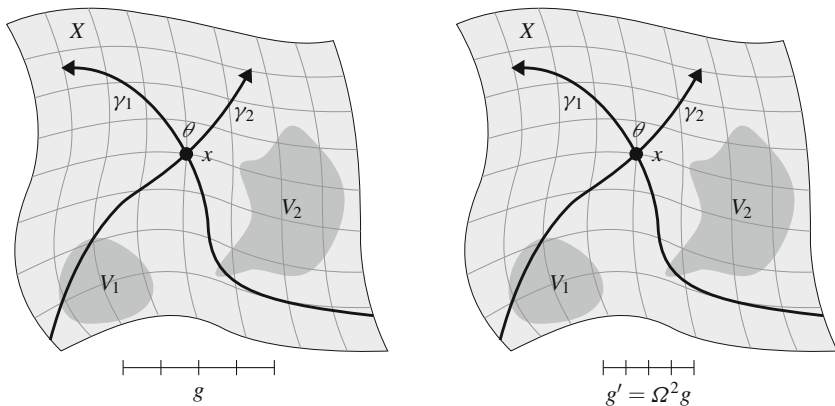


Figure 2.5.1. “Angle data” and “scale data” in relativistic spacetime; multiplying the metric by a constant factor changes “scales,” but does not affect “angles”.

number Ω . To be precise, g' is shorthand for the metric defined by the formula $g'_x(v, w) = \Omega^2 g_x(v, w)$ at each point x in X . In this particular example, Ω is chosen to be greater than 1; later, it will be allowed to assume any positive magnitude, and will also be allowed to vary over X . In the present case, the new metric g' will yield larger “scale” measurements for regions in X than g does; this is illustrated heuristically in the figure, by “drawing the tick marks on the measuring stick representing g' as closer together than the tick marks on the measuring stick representing g .” However, both metrics yield the *same* measurement for the “angle” between the curves γ_1 and γ_2 . Hence, “angle data” and “scale data” are separate, in the sense that two metrics may encode exactly the same “angle data,” even if they encode totally different “scale data.”

Conformal equivalence. The two metrics g and g' illustrated in Figure 2.5.1 provide a simple example of *conformally equivalent* metrics. The factor Ω , which distinguishes the two metrics, is called a constant *conformal factor*. The qualitative meaning of “conformal” is “same shape,” just as one would expect on etymological grounds. Since the conformal factor Ω is constant in this particular example, it is clear that g and g' measure any subset of X to be the “same shape, but different sizes;” examples of such subsets include the images of the curves γ_1 and γ_2 , and the shaded regions labeled V_1 and V_2 . The general concept of conformal equivalence is less restrictive: two pseudo-Riemannian metrics g and g' on a smooth manifold X are called *conformally equivalent* if they measure the “same shapes” in an *infinitesimal* sense, i.e., if “sufficiently small regions are arbitrarily close to being the same shape” with respect to the two metrics. This notion is made precise in Definition 2.5.2 below.

In Riemannian geometry, the idea that two metrics g and g' on a manifold X measure the same shapes in an infinitesimal sense may be elegantly re-expressed by the simple statement that g and g' measure the same *angles* between pairs of tangent vectors in the tangent spaces $T_x X$ of X . For this reason, “angle preservation” is often

the central motivating concept offered when conformal structure is introduced in elementary settings; for example, in single-variable complex analysis. In *pseudo-Riemannian* geometry, the same intuition remains useful, but true angles do not play the same role in the case of mixed signature, since in this case a metric on X does not define true inner products on its tangent spaces. In this context, it is simpler to describe conformal equivalence of metrics in terms of *how such metrics may differ*; namely, with regard to scale measurements. This is accomplished by simply allowing the constant conformal factor Ω in the previous example to vary smoothly over X .

Definition 2.5.2. *Two pseudo-Riemannian metrics g and g' on a smooth manifold X are called **conformally equivalent** if there exists a smooth positive function*

$$\Omega : X \rightarrow \mathbb{R},$$

*called the **conformal factor**, such that $g'_x(v, w) = \Omega(x)^2 g_x(v, w)$ for every point $x \in X$, and every pair of tangent vectors $v, w \in T_x X$.*

A **conformal geometry** on a smooth manifold X is an equivalence class of conformally equivalent pseudo-Riemannian metrics on X . As a branch of mathematics, conformal geometry is concerned with “scale-independent” properties of manifolds. Conformal geometry is important in many physics-related contexts besides general relativity. Perhaps the most famous of these, already mentioned in Section 2.4, is Maldacena’s AdS/CFT correspondence in string theory, where “AdS” stands for “anti-de Sitter,” and “CFT” stands for “conformal field theory.” A less conventional application is Julian Barbour’s *shape dynamics* [BA12], which studies the “evolution of three-dimensional conformal spatial geometries.” More recently, Penrose’s *conformal cyclic cosmology* [PE10] examines models of the universe in which the contribution of scale data is “transient” in a limiting sense, allowing cosmological epochs to be “stitched together,” despite the “initial smallness of Big-Bang type scenarios,” and the “terminal largeness of expanding spacetimes.”

“Scaling the manifold instead of the metric.” A familiar scientific fable features an observer who wakes up one morning to find that everything, except for all the measuring sticks, has increased in size during the night. The question then becomes whether “the world has really grown,” or whether “the measuring sticks have shrunk.” Silly as this scenario may seem, reputable physicists have actually worked on very similar ideas; for example, the question of whether or not there is any important distinction between the conventional wisdom that *spacetime is expanding*, and the alternative hypothesis that *its material content shrinking*. The answers to such questions depend on a number of factors, but one of the most obvious of these is the actual *nature* of fundamental spacetime structure. It is useful to consider this question in the specific context of conformal structure, since changes of scale are “allowed” in this setting. In trading the metric g on X for the conformally equivalent metric $g' = \Omega^2 g$, as illustrated in Figure 2.5.1, one is not “doing anything” to the underlying manifold X , so in this case it is clear that “the measuring sticks have shrunk.” Intuitively, it is easy to imagine the alternative scenario, in which “ X grows;” this scenario is illustrated

in Figure 2.5.3. In attempting to make this idea precise, however, an obvious problem arises: how can one “scale the manifold instead of the metric,” since a metric is required to determine “sizes in X ” in the first place?

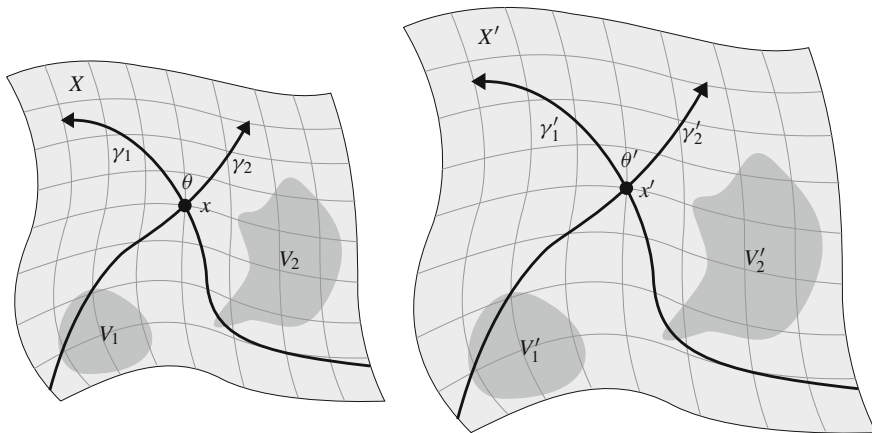


Figure 2.5.3. “Scaling the manifold instead of the metric” makes sense only if scale data is available from some other source.

The only possible solution to this problem is that *one must obtain scale data about X from some other source*, and this is exactly what I have done in the figure. Here, for once, it is instructive to think for a moment about the figure itself, temporarily forgetting about what it is supposed to represent abstractly. The illustration of X looks like a manifold; or, more precisely, like “part of a manifold,” *embedded in a higher-dimensional space, which is itself equipped with a metric*. This metric provides a “natural way to measure X .” Given this setup, it is easy to “copy,” and “dilate” X to yield another manifold, called X' , embedded in the same ambient space. The new manifold X' may then be measured using the same metric, and it is clear that in this case “the world has grown,” while “the measuring sticks have stayed the same size.” Of course, in pseudo-Riemannian geometry, there is generally no ambient space to appeal to, and therefore no “natural way to measure X ,” instead, one must *choose* a metric. In the discrete causal context, however, there *is* a natural source of scale data for classical histories; namely, discrete causal structure itself. This is one of the principal reasons why the metric recovery theorems provide “enough evidence” to motivate the classical causal metric hypothesis (CCMH) in the discrete setting.

Smooth conformal isometries. Returning to Figure 2.5.3, what is represented is a “pair of spacetimes” X and X' embedded in an ambient space, whose metrics g and g' are induced by a choice of metric on this space. In this context, there exist natural diffeomorphisms between X and X' , which may be called *dilation* and *contraction maps*. In particular, the dilation map sends x to x' , and the contraction map sends x' to x . These maps are simple examples of *smooth conformal isometries*,

i.e., isomorphisms between X and X' in the smooth conformal category. It is worth emphasizing that referring to a map f as a “conformal isometry” does *not* mean that f is “an isometry that is conformal,” since isometries are *always* conformal by definition. Rather, it means that f is an “isometry up to conformal equivalence.” This is made precise by the following definition:

Definition 2.5.4. A **smooth conformal isometry** $f : (X, g) \rightarrow (X', g')$, between two pseudo-Riemannian manifolds (X, g) and (X', g') , is a diffeomorphism $f : X \rightarrow X'$, such that the pulled-back metric f^*g' on X is conformally equivalent to the metric g on X .

The reason for the qualifier “smooth” in the phrase “smooth conformal isometry” is that it is possible to study conformal properties of objects in more general categories, as already mentioned in Section 2.3. A *smooth conformal morphism* $f : (X, g) \rightarrow (X', g')$ is defined by demoting “diffeomorphism” to “smooth map” in Definition 2.5.4. Since conformal equivalence of metrics “preserves shapes” only infinitesimally, smooth conformal morphisms and smooth conformal isometries generally do *not* “preserve shapes” at a finite level, and they obviously do not preserve volumes, lengths, and other related measurements. Figure 2.5.5 illustrates two spacetimes related by a smooth conformal isometry, which are *not* “the same shape” macroscopically. The checkerboard pattern in the figure is included to show how corresponding regions “approach the same shape as they shrink in size;” in particular, the individual black and gray regions in the right-hand diagram are “much closer to being squares” than the region itself.

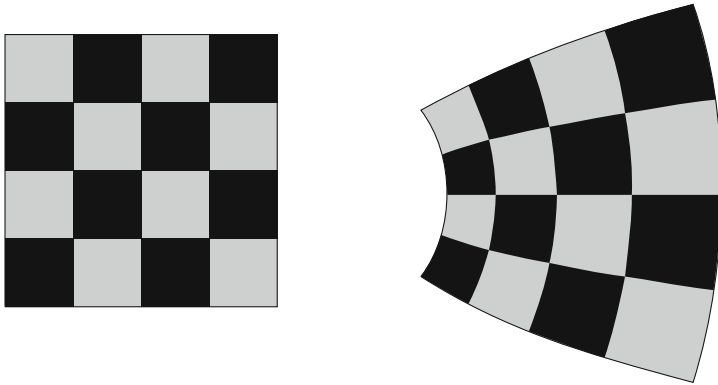


Figure 2.5.5. Spacetimes related by a smooth conformal isometry.

Passive viewpoint versus active viewpoint. The specific distinction between *conformal equivalence of metrics* and *conformal morphisms between manifolds* provides an excellent illustration of the more general difference between the passive and active viewpoints regarding structural comparisons, already mentioned in Section 2.3

in the context of category theory. Exchanging one metric for another involves the passive viewpoint, since one begins with an underlying structure, which is “left alone,” while alternative additional structures are superimposed on it. Such an exchange of metrics is mathematically analogous to changing “coordinate systems” or “frames of reference,” but is much more drastic, at least in the relativistic context. This is because different metrics are generally assumed to encode different physics,¹⁸ while different frames of reference merely represent different points of view regarding the same physics. Conformal morphisms, by contrast, involve the active viewpoint, since they map each element of one manifold to an image element in another manifold. Of course, nothing *physically* passive or active is involved in either case; but the choice of terminology offers a convenient way of describing the difference in viewpoint.

Subtle relationships exist between the active and passive viewpoints in the context of conformal structure. For example, if g and $g' = \Omega^2 g$ are conformally equivalent metrics on an underlying smooth manifold X , then the identity map $\text{Id} : X \rightarrow X$ may be viewed as a map between two *different* pseudo-Riemannian manifolds (X, g) and (X, g') . From this viewpoint, the map “Id” is no longer the identity if $g \neq g'$; i.e., it is not the identity morphism in the pseudo-Riemannian category, because it is a map between two different objects in the category. In fact, it is not a morphism in this category at all, because it fails to preserve metric structure. However, it *is* a conformal isometry relating the two metrics g and g' , because $\text{Id}^* g' = g' = \Omega^2 g$. Given such a pair of conformally equivalent metrics on X , there generally exist many different self-diffeomorphisms $f : X \rightarrow X$ relating g and g' via the same conformal factor; for example, the underlying diffeomorphism of any smooth self-isometry f of (X, g) relates g to itself in the trivial way: $f^* g = g = 1^2 g$.

One may also ask the question of when a nontrivial self-diffeomorphism $f : X \rightarrow X$, viewed as a self-map on a *fixed* pseudo-Riemannian manifold (X, g) , is a conformal isometry relating g to *another* previously-chosen metric $g' = \Omega^2 g$ on the underlying smooth manifold X , i.e., for which f it is true that $f^* g = g'$. If one begins with a smooth conformal isometry f , then by definition one obtains a pair of conformally equivalent metrics, but the conformal factor relating them is determined by the details of f , and cannot be specified beforehand. If, on the other hand, one first *chooses* a pair of conformally equivalent metrics g and g' related by a particular conformal factor Ω^2 , then there may or may not exist any self-diffeomorphism $f : X \rightarrow X$ such that $f^* g = g'$.¹⁹ This provides a preliminary hint that the passive viewpoint has information to offer that is unavailable from the active viewpoint.

¹⁸Here I am referring to the many attempts, from Weyl to CFT’s to conformal cyclic cosmology, to extend the relativity principle to include *conformal invariance*. In such a theory, an appropriate change of conformally equivalent metrics would *not* affect the physics. Note that under the passive viewpoint, the question of whether or not “different metrics encode different physics” does *not* involve the active idea of “re-associating physical data with different points in a manifold,” discussed in Section 2.3 in the context of diffeomorphism invariance. Here, the question is merely, “if the metric near x is changed, does the physics near the event represented by x change?”

¹⁹See Wald [WA84], Appendix D.

2.6 Causal Structure

Relativistic causal structure and the classical causal metric hypothesis. What the metric recovery theorems demonstrate is that *the causal structure of relativistic spacetime determines the corresponding metric structure up to smooth conformal isometry*. Informally, this means that information about cause and effect reveals everything that there is to know about relativistic spacetime geometry, *except* for scale data, i.e., except for the choice of conformal factor. As mentioned in Section 2.3, if classical spacetime *really is* precisely represented by a pseudo-Riemannian manifold, then the metric recovery theorems are not quite sufficient to support the classical causal metric hypothesis (CCMH), since there is no way to obtain the “missing conformal factor” without turning to some auxiliary, “non-causal” source. Hence, if general relativity is *absolutely correct* in how it models classical spacetime, then the classical causal metric hypothesis is *wrong*, and the metric recovery theorems explain exactly why, and to what extent. This does not necessarily mean that the physical existence of continuum-based structure *in general* would doom the causal metric hypothesis, since one could conceive of continuum-based spacetime models other than pseudo-Riemannian manifolds. For example, *domain theory* provides a context in which “continuous, pure causal” structure makes sense. On the other hand, if classical spacetime is discrete, and merely *looks* like a pseudo-Riemannian manifold at large scales, then a natural scale comes for free: the scale of the fundamental elements and relations.²⁰ This is what the founders of causal set theory realized around 1980.

Directions of curves in relativistic spacetime. What, precisely, is meant by “*the causal structure of relativistic spacetime?*” The reader is no doubt well-aware that relativity “forbids superluminal communication,” and this means that influence may travel only along certain curves²¹ in spacetime, called *causal curves*. These curves are determined by the metric g on a relativistic spacetime manifold X , which supplies information about “what direction a differentiable curve is pointing” at each point along the curve. At a fixed point x in X , these various “directions” may be partitioned into three classes, usually called *timelike*, *null*, and *spacelike*. Timelike and null directions may be further subdivided into classes of *past* and *future* directions. From a physical standpoint, future timelike or null directions are directions in which “causal influence may propagate,” while past timelike or null directions are directions from which “causal influence may arrive.” Hence, these directions have absolute physical significance, and their identification does not depend on a choice of reference frame. For spacelike directions, however, the distinction between “past” or “future” *does* depend on the frame of reference, as elaborated below. If X admits a consistent, continuously-varying designation of “past” and “future” for timelike and null directions, then it is called *time-orientable*, and most references on general

²⁰As explained in Section 2.1, there may be more than one reasonable way to define “emergent volume” in terms of this fundamental scale data.

²¹Here I am really referring to the *images* of such curves; the distinction between a curve and its image is discussed in more detail below.

relativity eliminate non-time-orientable spacetime manifolds from consideration at the outset, on basic physical grounds. The possibility of non-time-orientability is one of many “continuum-related pathologies” that discrete causal theory avoids entirely, since there is no need for anything to vary continuously.

The “directions” at a point x in a relativistic spacetime manifold X may be represented precisely by tangent vectors to X at x . A tangent vector v at x in X is called *timelike* if $g_x(v, v) < 0$, *null* if $g_x(v, v) = 0$, and *spacelike* if $g_x(v, v) > 0$. These designations make sense for any pseudo-Riemannian manifold of Lorentz signature. In the special case in which X is Minkowski spacetime \mathbb{R}^{3+1} , X is isomorphic to $T_x X$, and each nonzero tangent vector v at x “actually points to another element y of X ,” in an obvious sense. In this case, g is the Minkowski metric η , and the Minkowski *spacetime interval*, or more precisely, “squared interval,” between x and y , is just $\eta_x(v, v)$. The events x and y are called *timelike separated* if $\eta_x(v, v) < 0$, *null separated* if $\eta_x(v, v) = 0$, and *spacelike separated* if $\eta_x(v, v) > 0$. In the general case of a curved spacetime manifold X , the corresponding relationships between pairs of events x and y are described in terms of curves from x to y , as explained below.

In relativistic kinematics, a future timelike direction is a “permissible direction” for the motion of a massive material object, such as an electron. A future null direction is a “permissible direction” for the propagation of electromagnetic radiation, or any other form of energy involving massless particles. In particular, future null directions at a point x in a relativistic spacetime manifold X may be viewed as the possible directions of light rays emanating from x , and these determine the “future light cone,” or *future null cone*, of x in the tangent space $T_x X$. Spacelike directions are “forbidden directions” for all forms of influence. Due to the relativity of simultaneity, a spacelike direction that points toward the future in some frames of reference will point toward the past in other frames. Timelike and null directions do not suffer from this ambiguity. The left-hand diagram in Figure 2.6.1 illustrates timelike, null, and spacelike future directions at a point x in X , with respect to a particular frame of reference, represented by the curved “coordinate lines.” Of course, the timelike and null directions illustrated here retain their future orientation in any frame of reference.

Smooth causal curves. The “directional information” associated with tangent vectors in a relativistic spacetime manifold X may be used to classify special families of curves in X . As in the case of tangent vectors themselves, the resulting definitions make sense for any pseudo-Riemannian manifold of Lorentz signature. A smooth curve passing through $x \in X$ is called *timelike at x* if its tangent vector at x is timelike. It is called *globally timelike*, or just *timelike*, if it is timelike at each of its points. A pair of distinct events x and y in X are called *timelike separated* if they are connected by a smooth timelike curve. *Null* and *spacelike* smooth curves, and *null* and *spacelike separation* of pairs of distinct events, are defined in an analogous manner. A smooth curve is called **causal** if it is either timelike or null at each of its points. The right-hand diagram in Figure 2.6.1 illustrates timelike, null, and spacelike future-directed smooth curves at a point x in X , with respect to the same frame

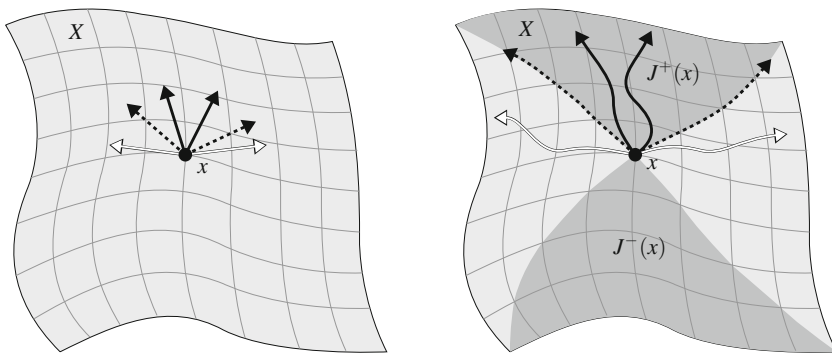


Figure 2.6.1. Timelike (black), null (dashed), and spacelike (white) future directions in a given frame of reference; corresponding timelike, null, and spacelike smooth curves.

of reference illustrated in the left-hand diagram. The warped, dark gray “triangular” regions labeled $J^-(x)$ and $J^+(x)$ in the figure represent the *causal past* and *causal future* of x , respectively; these are discussed more thoroughly below. The definitions of timelike, null, and spacelike curves given here do not require actual smoothness in order to make sense; these curves need only possess unique tangent vectors at each of their points.

A few mathematical details are worth mentioning here for the purpose of clarity. In the context of general relativity, a *curve* in a spacetime manifold X is usually defined to be a *map* from an interval in \mathbb{R} into X , often with additional conditions imposed, such as smoothness, or non-vanishing of the tangent vector along the curve. However, in the present context, one is principally interested in the properties of the *image* of such a map, rather than the details of the map itself. For example, “reparameterization” of such a map generally does not change its physical interpretation. In particular, the elementary-calculus heuristic of “moving at different speeds along the same curve” is irrelevant for a map into spacetime, whose image encodes all physical motion-related quantities. This means that very large classes of physically equivalent curves lurk beneath every curve-related notion in relativity. In particular, different curves sharing the same image in X will generally have different tangent vectors at a given point in X , so one might a priori worry about whether or not the notions of timelike, null, and spacelike curves, defined in terms of these tangent vectors, are actually physically meaningful. The reason why these particular notions *are* meaningful is because only the *sign* of the tangent vector, and not its magnitude, is involved in defining them. It is therefore common, and often innocuous, to refer to such curves and their images interchangeably, and many instances of such language appear throughout the remainder of the book. However, the presence of such equivalence classes must be kept in mind in more general settings. For example, the *paths* involved in Feynman’s path summation approach to ordinary quantum theory are large equivalence classes of curves in a spacetime manifold, sharing a common

image.²² Much like the issue of whether or not to associate physical events with specific elements in a pseudo-Riemannian manifold, which so troubled Einstein during the development of general relativity, the uncomfortable relativistic necessity of dealing with large equivalence classes of curves disappears in discrete causal theory. In this context, paths are often, though not always, represented by *individual morphisms*. These details are elaborated in Section 5.9.

Continuous causal curves. The definitions of timelike, null, and spacelike curves may be generalized further, to classes of *continuous* curves. In certain circumstances, this generalization makes an important difference to the physical significance of the classes of curves under consideration.²³ In particular, strictly stronger metric recovery results may be proven if one chooses to describe causal structure in terms of continuous causal curves, rather than restricting attention to smooth causal curves. This is the reason for the distinction between *causal morphisms* and *enhanced causal morphisms* of relativistic spacetime manifolds, appearing in Definition 2.6.3 below. The definitions of timelike, null, and spacelike curves in the continuous context are slightly subtle. For example, a continuous curve is called *timelike* if each of its points x possesses a *convex normal open neighborhood* U_x , such that any two points w and y on the curve in U_x are connected by a smooth timelike curve “in the proper order.”²⁴ *Null* and *spacelike* continuous curves are defined in an analogous manner. A continuous curve is called **causal** if it is everywhere timelike or null. The left-hand diagram in Figure 2.6.2 illustrates a continuous causal curve, with two points w and y in a suitable neighborhood U_x of a point x on the curve connected by a smooth causal curve.

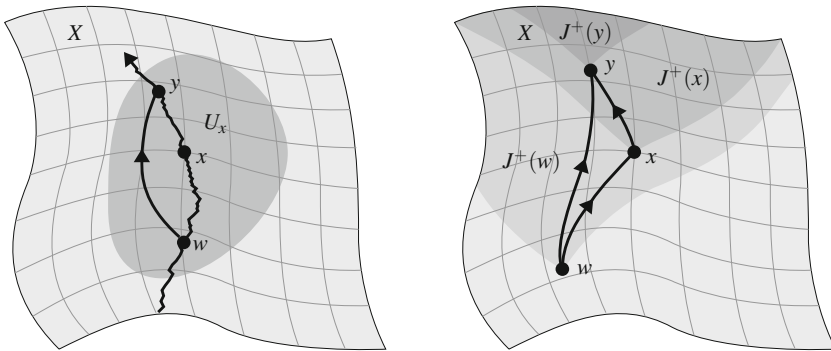


Figure 2.6.2. A continuous causal curve; transitivity of the relativistic causal relation.

²²For examples of the distinction between “curves” and “paths” in general relativity, see Rovelli [RO04], pp. xxii–xxiii, Thiemann [TH07], p. 164, Hawking and Ellis [HE73], p. 15, or Wald [WA84], p. 17.

²³See Malament, p. 1400.

²⁴See Wald [WA84], p. 193, for details.

Relativistic causal relation; causal pasts and futures. It is very useful to define a binary relation $<_{\text{GR}}$ on a relativistic spacetime manifold X , called the **relativistic causal relation**, where $x <_{\text{GR}} y$ if and only if there exists a nontrivial²⁵ smooth causal curve from x to y in X . In other words, $x <_{\text{GR}} y$ if and only if x and y are timelike or null separated. The subscript “GR” in the expression $<_{\text{GR}}$ stands for “general relativity,” and is included in order to avoid confusion with other binary relations appearing in this book, particularly the *causal relation* $<$ on a directed set in discrete causal theory, introduced in Section 3.10. As usual, however, the definition of $<_{\text{GR}}$ makes sense for any pseudo-Riemannian manifold of Lorentz signature. The set $J^-(x) := \{w \in X \mid w <_{\text{GR}} x\}$ is called the **causal past** of x , while the set $J^+(x) := \{y \in X \mid x <_{\text{GR}} y\}$ is called the **causal future** of x . These sets may be viewed as relativistic prototypes of the more general *domains of influence* introduced in Section 3.7, which are subsets or subobjects of a directed set D , defined to encode information about influences between pairs of events represented by elements of D . In the relativistic context, $J^-(x)$ is interpreted as the set of all events in X which “could conceivably” influence x , and $J^+(x)$ is interpreted as the set of all events in X which “could conceivably” be influenced by x . The union $J(x) = J^-(x) \cup J^+(x)$ is called the **total domain of influence** of x .

The causal past $J^-(x)$ and causal future $J^+(x)$ of an element x in a relativistic spacetime manifold X were illustrated as the warped, dark gray “triangular” regions appearing in the right-hand diagram in Figure 2.6.1 above. In the right-hand diagram in Figure 2.6.2, the progressively darker-shaded regions represent the causal futures of the events labeled w , x , and y . The boundaries of these regions generalize the future “light cones,” or *null cones*, of events in Minkowski spacetime \mathbb{R}^{3+1} . Two rather trivial points are worth mentioning here for the sake of clarity. First, the reason why these regions appear triangular, rather than cone-shaped, in these particular illustrations, is because only two of the dimensions involved are actually shown in the diagrams. Second, the term “light cone” is popular in the relativistic context, but is not ideal in *general relativity*. In this setting, the actual *cones* involved exist only in the tangent spaces $T_x X$ of a relativistic spacetime manifold X , and not in X itself, while rays of light follow curved geodesics in X , determined by Einstein’s equation (1.3.1). Hence, one must choose between speaking about a “light cone” in $T_x X$, which does not really describe the propagation of light, or a “light cone” in X , which is not really a cone. Penrose offers the reasonable suggestion to simply use the more precise term *null cone* to describe the desired object in $T_x X$; one may then describe the more general corresponding structures in X itself as *pasts* and *futures*, eliminating the term “light cone” entirely. I mostly follow Penrose’s convention in this book.

The relativistic causal relation $<_{\text{GR}}$ is *transitive*, which means that if $x <_{\text{GR}} y$ and $y <_{\text{GR}} z$, then $x <_{\text{GR}} z$. For relativistic spacetime manifolds satisfying the *causal condition*, defined in Section 2.7, $<_{\text{GR}}$ is also *irreflexive*, which means that $x \not<_{\text{GR}} x$. At a conceptual level, transitivity of $<_{\text{GR}}$ encodes the “common sense” that “if x influences y , and y influences z , then x influences z .” The choice to abstain from

²⁵“Nontrivial” means “nonconstant;” see the discussion of the *causal condition* in Section 2.7 for more details.

imposing transitivity on the more general *causal relation* introduced in Section 3.10 does not contradict this common sense, but merely recognizes the fact that discrete structure demands a more fundamental notion of *direct* causation, encoded by \prec , while “possibly indirect” causation is encoded by a transitive relation \prec_{tr} , generated by \prec . Irreflexivity of \prec_{GR} means that “ x does not influence itself;” however, if there exists a nontrivial *closed causal curve* in X , beginning and terminating at x , then transitivity implies that x *does* influence itself. The causal condition rules out the existence of such curves, which is why irreflexivity of \prec_{GR} holds only for spacetime manifolds satisfying this condition. Conveniently, it turns out that the actual identity of $J^+(x)$ and $J^-(x)$, as subsets of X , is unaffected if “smooth causal curve” is replaced by “continuous causal curve” in the definition. This follows from the transitivity of the relativistic causal relation, together with an easy compactness argument. By replacing causal curves with timelike curves in the definition of the relativistic causal relation, one may define a corresponding relativistic *chronological relation*, which I denote by \prec_{GR} . In particular, the set $I^-(x) := \{w \in X \mid w \prec_{\text{GR}} x\}$ is called the *chronological past* of x , while the set $I^+(x) := \{y \in X \mid x \prec_{\text{GR}} y\}$ is called the *chronological future* of x . Finally, one may define a *horismos relation*, sometimes denoted by $x \rightarrow y$, by using only null curves in the same setting.²⁶

Causal morphisms and “enhanced causal morphisms.” The relativistic causal relation \prec_{GR} on a relativistic spacetime manifold X is what is usually meant by the *causal structure of relativistic spacetime*. It is curious, however, that incorporating information about *continuous* causal curves, as opposed to merely *smooth* causal curves, yields stronger metric recovery results. In discrete causal theory, there are no such distinctions, and hence no ambiguities about how structure-preserving maps, i.e., morphisms, should be defined. To deal with this particular continuum-induced complication, I give *two* definitions of causal morphisms: the first following the usual relativistic conventions, and the second included to account for the difference between smooth curves and continuous curves in this context.

Definition 2.6.3. *Let $f : X \rightarrow X'$ be a map between two pseudo-Riemannian manifolds of Lorentz signature, with relativistic causal relations \prec_{GR} and \prec'_{GR} , respectively.*

1. *The map f is called a **causal morphism** if it preserves relativistic causal relations, i.e., if $f(x) \prec'_{\text{GR}} f(y)$ whenever $x \prec_{\text{GR}} y$. It is called a **causal isomorphism** if it possesses an inverse that is also a causal morphism.*
2. *The map f is called a **enhanced causal morphism** if it preserves future-directed continuous causal curves. It is called an **enhanced causal isomorphism** if it possesses an inverse that is also an enhanced causal morphism.*

The term “enhanced causal morphism” is awkward; the alternative term “strong causal morphism” is a priori more attractive. However, the former term is chosen here to avoid confusion with the *strongly causal condition* on relativistic spacetime manifolds, discussed in Section 2.7, which is ubiquitous in the early-modern relativity

²⁶See Malament p. 1400 for details.

literature. The definitions given here are stronger than necessary to prove the version of metric recovery stated in Theorem 2.8.1; in particular, Malament [MA77] works mostly in terms of maps f and f^{-1} that are assumed only to preserve certain classes of *timelike* curves. The necessary properties for causal curves in general are then *proven*, using the assumptions regarding timelike curves, together with the topological structures of the manifolds involved. However, these details are not essential to the present discussion. Causal morphisms generalize in a natural way to the context of directed sets, but enhanced causal morphisms do not, because there is generally no concept of continuity in this setting. Note that this distinction cannot be eliminated by defining a new binary relation on a relativistic spacetime manifold X in terms of *continuous* causal curves, since one merely recovers the usual relativistic causal relation $<_{\text{GR}}$ in this manner. This, of course, is just a pointwise statement of the fact that the causal pasts and futures $J^+(x)$ and $J^-(x)$ of an event x in X do not depend on whether one uses smooth curves or continuous curves to define them. This situation foreshadows the conclusion that *something more than a simple binary relation is necessary to specify geometry in the absence of natural scale data*.

2.7 Causality Conditions

Avoiding “causality violations.” The degree to which the metric structure of a relativistic spacetime manifold X may be recovered from its causal structure depends, informally, on “how close X comes to violating causality.” In this context, “violating causality” means that X includes events that influence their own causes, and hence, that indirectly influence themselves. *Causality conditions* are technical conditions that give precise meaning to the words “how close,” in this qualitative description. As discussed below, the potential consistency issues posed by “causality-violating” relativistic spacetime manifolds may be completely avoided in discrete causal theory, by working in a perfectly background independent setting. However, causality conditions remain important in this context, due to their role in metric recovery.

Overview of causality conditions. Seven causality conditions that feature prominently in general relativity are the *chronological*, *causal*, *past or future distinguishing*, *past and future distinguishing*, *strongly causal*, *stably causal*, and *globally hyperbolic* conditions. Figure 2.7.1, which is an elaboration of one of Malament’s [MA77] diagrams, lists these conditions in ascending order of restrictiveness. For example, every strongly causal spacetime is past and future distinguishing, but the converse is false. Conditions possessing natural analogues in discrete causal theory appear in bold font in the figure. From a modern viewpoint, the past and future distinguishing condition is the condition that determines whether *enhanced* causal isomorphisms, or merely causal isomorphisms, are necessary in the hypotheses for metric recovery. This is the reason for the dashed line in the figure, which separates conditions “restrictive enough” to render enhanced causal isomorphisms unnecessary from conditions which are “too weak.” One of Malament’s achievements was to eliminate the use

of the “strongly causal” hypothesis from Hawking’s earlier proof. The remainder of this section examines these seven conditions. Due to the relationships among them, it is convenient to discuss them in a somewhat different order: the chronological and causal conditions are discussed first, followed by the stably causal condition, the past and/or future distinguishing conditions, the strongly causal condition, and finally, the globally hyperbolic condition.

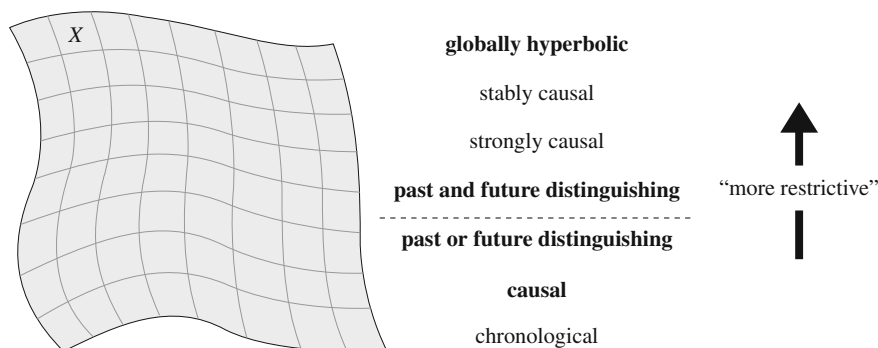


Figure 2.7.1. Seven causality conditions on a relativistic spacetime manifold X .

Closed timelike curves; chronological condition. The “worst” possible type of behavior, from the viewpoint of “causality violation,” is the existence of *closed timelike curves*, like the one illustrated in the left-hand diagram in Figure 2.7.2. A relativistic spacetime manifold sufficiently “well-behaved” not to contain such curves is called *chronological*. Since relativity permits the transport of a material body in any future timelike direction, an observer could theoretically travel to his or her own past along such a curve. From a naïve viewpoint, this possibility raises serious consistency issues. In particular, familiar “causality paradoxes,” such as the *grandfather paradox*, originate from this scenario. It is likely a waste of time to devote serious consideration to such “paradoxes” in their own right, since their actual significance seems to be mostly restricted to illustrating potential issues arising from a lack of perfect background independence in general relativity. Indeed, the common source of such paradoxes is a potential clash between two different types of structure: the relativistic spacetime manifold on which a material body could conceivably “move around and return to where it started,” and the material body itself. The strong interpretation of the causal metric hypothesis (CMH) eliminates such issues once and for all, by admitting *only one type of basic structure*; namely, causal structure, of which “spacetime” and “material bodies” are viewed as different manifestations. It does not rule out the possibility of discrete causal analogues of closed timelike curves, i.e., *cycles* such as the one illustrated in the right-hand diagram in Figure 2.7.2, but it does guarantee their physical consistency, should they exist. In particular, it avoids the potential conflicts involved in traversing a cycle, by simply disallowing any extrinsic entity that *could* traverse it. Of course, most versions of discrete causal

theory do not distinguish precisely between analogues of timelike and null curves, so it is generally not obvious which cycles in a directed set should be regarded as “timelike.” For this reason, the chronological condition has no simple analogue in discrete causal theory.

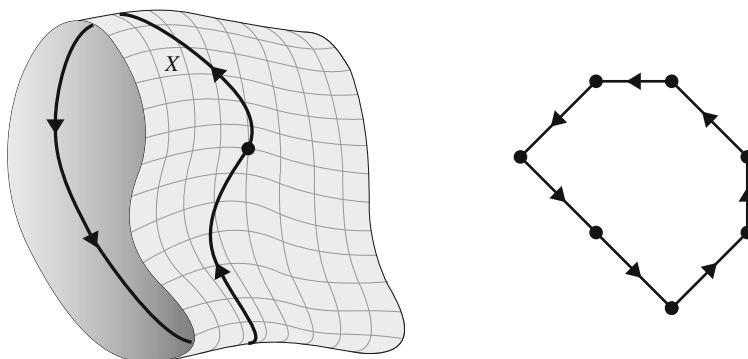


Figure 2.7.2. A closed timelike curve in relativistic spacetime; a cycle in a directed set.

Closed causal curves; causal condition. A slightly more restrictive causality condition than the chronological condition is the *causal condition*, which rules out not just closed timelike curves, but *all* nontrivial closed causal curves, including null curves. Any pseudo-Riemannian manifold X admits trivial “constant curves,” mapping an entire interval in \mathbb{R} to a single point x in X , and these curves are technically “null,” because their tangent vectors vanish. Such curves are *not* interpreted as encoding “self-causation.” Throughout the remainder of the book, null curves are assumed to be nontrivial unless stated otherwise. The principal qualitative difference between a closed timelike curve and a closed null curve is that relativity permits the latter to convey only *information*, and not material bodies. In the absence of perfect background independence, however, closed null curves are “nearly as bad” as closed timelike curves from a consistency standpoint. For example, one may imagine sending instructions “back in time” along a closed null curve, detailing how to construct the very device one is using to send the instructions. This creates obvious problems if one distinguishes “spacetime” from “information,” since once again two types of structure are pitted against each other.

By contrast, discrete causal analogues of non-causal spacetimes, viewed according to the strong interpretation of the causal metric hypothesis (CMH), are immune to consistency issues involving cycles, since they possess only one type of structure. If such cycles *do* exist, they merely represent “part of what the causal structure is like,” even if the resulting behavior is alien to ordinary experience. However, it is worth noting that the class of relativistic spacetime manifolds admitting the “best” metric recovery results, i.e., results that do *not* require the use of enhanced causal isomorphisms, is a subclass of the class of causal spacetimes; namely, the class of *past and future distinguishing* spacetimes. In other words, it is technically more difficult

to recover metric structure from causal structure for spacetimes that fail to satisfy the causal condition. It is possible to interpret this fact as a hint that acyclicity *should* be taken as an axiom of discrete causal theory. In my own opinion, however, the relativistic evidence is not overwhelming one way or the other; for instance, closed causal curves appear in rather generic relativistic situations, such as the *Kerr black hole*. An important characteristic of relativistic spacetime manifolds satisfying the causal condition is that the relativistic causal relation \prec_{GR} on such a manifold is in fact a *strict partial order*, i.e., an acyclic transitive binary relation. For this reason, *causal sets may be viewed as discrete causal analogues of causal spacetimes*, since causal set theory restricts its consideration of directed structure to the order-theoretic paradigm.

Stability issues; stably causal condition. Relativistic spacetime manifolds satisfying the causal condition may be viewed as “causally well-behaved,” since they avoid, by definition, causality-based conflicts between “spacetime” and auxiliary “matter-energy content.” However, the causal condition is often inadequate from a practical perspective. In particular, an arbitrarily small perturbation of the metric in a causal spacetime can produce closed causal curves. Issues of this nature arise in almost any conceivable continuum-based theory, as part of the cost to be paid for the availability of convenient interpolation and limiting procedures. To prove certain desirable results, it is sometimes necessary to impose more restrictive conditions that “bound systems away from bad behavior,” instead of merely ruling out the bad behavior itself. The *stably causal* condition, appearing second on the list in Figure 2.7.1, accomplishes this by explicitly requiring that a particular type of finite perturbation of the metric preserves the chronological condition.²⁷ In the context of metric recovery, however, the stably causal condition is unwieldy. Moreover, the condition does not possess a natural discrete causal analogue.

Past and/or future distinguishing conditions. Less-restrictive conditions, called past and/or future distinguishing conditions, are more useful in this setting. Informally, these conditions govern the extent to which a relativistic spacetime manifold X “separates pasts and futures of individual events.” In particular, X is called *past distinguishing* if distinct elements possess distinct causal pasts, and is called *future distinguishing* if distinct elements possess distinct causal futures. Conditions of this nature are easy to conceptualize by considering relationships among small finite sets of elements, and this reflects the fact that these conditions possess natural analogues in discrete causal theory. The directed sets depicted in the left-hand diagram in Figure 2.7.3 illustrate these conditions. The “diamond-shaped” set D is neither past nor future distinguishing, since the elements x_1 and x_2 both possess the same past and the same future. The set D' is future distinguishing but not past distinguishing, because the elements y_1 and y_2 both possess the same past. The set D'' is past and future distinguishing. For simplicity, I am ignoring the “sameness” of empty pasts and/or futures in these examples. It is worth noting that a “large random directed set” may easily fail to satisfy either or both of these conditions, due solely to small, local

²⁷See Wald [WA84] p. 198 for details.

“defects,” such as the existence of “diamond-shaped” subsets. Hence, one should be very reluctant to impose such conditions as *axioms* for discrete causal theory.

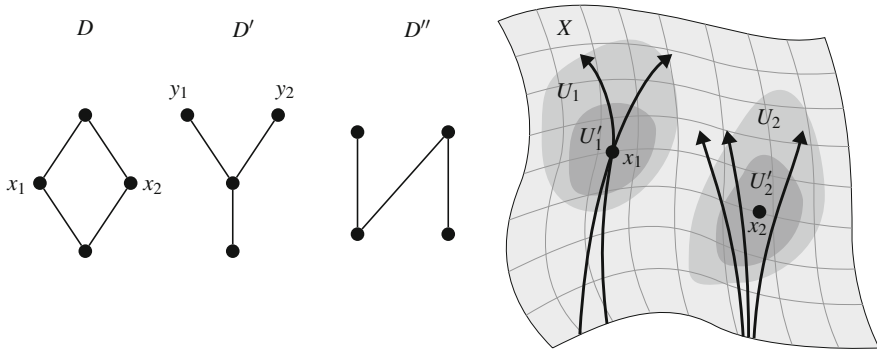


Figure 2.7.3. Directed sets illustrating past and/or future distinguishing conditions; relativistic versions of the future distinguishing condition and the strongly causal condition.

Returning to the relativistic context, it is convenient to reformulate past and/or future distinguishing conditions in *topological* terms. In particular, it turns out that a relativistic spacetime manifold X is future distinguishing if and only if for every event x in X , and every open set U containing x , there exists a “smaller” open set U' in U , containing x , such that no future-directed smooth timelike curve through x that leaves U' ever returns to it.²⁸ This situation is illustrated for the event x_1 in the right-hand diagram in Figure 2.7.3. The past distinguishing condition may be reformulated in an analogous manner, by replacing the word “future” with “past” in the reformulation of the future distinguishing condition.

Strongly causal condition. The topological reformulation of the past and/or future distinguishing conditions may be slightly modified to yield a stronger condition, called the strongly causal condition. A relativistic spacetime manifold X is called *strongly causal* if for every element x in X , and every open set U containing x , there exists a “smaller” open set U' in U , containing x , such that no future-directed smooth timelike curve leaving U' , *whether or not it passes through x* , ever returns to U' . Heuristically, this condition provides “slightly more room between pasts and futures” than the past and/or future distinguishing conditions. The strongly causal condition is illustrated for the event x_2 in the right-hand diagram in Figure 2.7.3. There is generally no natural analogue of this condition in discrete causal theory, and it plays a less-significant role even in relativity than it did a few generations ago. As mentioned above, Malament succeeded in removing this condition from the hypotheses of his version of metric recovery. More recently, Bernal and Sanchez [BS07] have expunged it from topological reformulations of the globally hyperbolic condition, discussed below.

²⁸See Malament p. 1400 for details.

Cauchy surfaces; globally hyperbolic condition. The globally hyperbolic condition is the most restrictive causality condition of the seven listed in Figure 2.7.1, and requires a bit of preliminary explanation before it may be properly introduced. First, the terminology originates from the study of *wave equations*, i.e., special *hyperbolic partial differential equations*, on continuum-based models of spacetime. In this general context, one of the variables is interpreted as representing time, whether in a Euclidean or Lorentzian fashion, and the differential equations under consideration are used to model *initial value problems*. The informal idea behind this approach, handed down from pre-relativistic physics, is that if one knows the values of certain quantities “everywhere in space, at a given instant in time,” then one can solve for the corresponding values at all later times. Since this description refers implicitly to *simultaneity*, which is not absolute in relativity, one must be precise about which subsets of relativistic spacetime manifolds are suitable representatives of the notion of “everywhere in space, at a given instant in time.” Such subsets are called *Cauchy surfaces*. The corresponding initial value problems are classified as special types of *Cauchy problems*, i.e., problems involving the solution of partial differential equations satisfying specified conditions on hypersurfaces. Since discrete causal analogues of Cauchy surfaces play a crucial role in this book, it is worthwhile to give a formal definition:

Definition 2.7.4. A *Cauchy surface* in a relativistic spacetime manifold X is a subset σ of X such that every inextendible causal curve in X intersects σ exactly once.

The term “inextendible causal curve” in the definition means what the terminology suggests: a causal curve that cannot be extended to yield a “longer” causal curve. Several different types of inextendible causal curves exist. One obvious type is *closed* causal curves, such as the closed causal curve illustrated in Figure 2.7.2. A few other types of inextendible causal curves are illustrated in the left-hand diagram in Figure 2.7.5. The white region represents a “hole” in the relativistic spacetime manifold W , with its boundary “stripped away.” Of course, it is not a hole in a physical sense, but a topological feature of W . Causal curves in W that “approach this hole,” such as the curves γ_2 , γ_3 , and γ_4 illustrated in the figure, cannot be extended. A “simpler” type of inextendible causal curve is one that “runs on forever in both directions,” such as the curve γ_1 .

A Cauchy surface in a relativistic spacetime manifold X is a special case of an *acausal subset* of X , which is defined to be a subset intersected *at most once* by any causal curve. In particular, no pair of events belonging to such a subset are causally related. The right-hand diagram in Figure 2.7.5 illustrates two acausal subsets of X , represented by thick horizontal curves. The lower, “broken” curve, labeled ρ , is *permeable*, in the sense that it has “gaps,” through which causal curves may pass without intersecting it. The upper curve, labeled σ , is a Cauchy surface, which is by definition *impermeable*; no causal curve may pass from its past to its future without intersecting it. From a modern information-theoretic viewpoint, a Cauchy surface “samples,” or “filters,” data flowing from its past to its future. A permeable acausal subset is “faulty” in this sense, because information may flow from its past to its future without being “sampled,” by following causal curves permeating the subset.

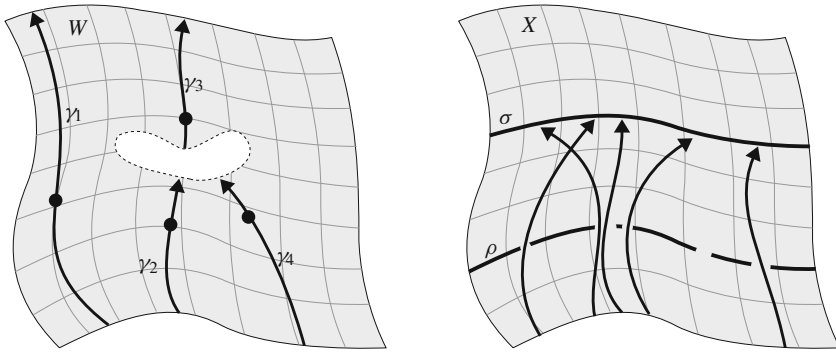


Figure 2.7.5. Inextensible causal curves; a permeable acausal subset ρ and a Cauchy surface σ .

This creates “predictability issues,” since data associated with a permeable acausal subset is insufficient to determine “what will happen” in the future of the subset. This is sometimes expressed by saying that the initial value problem for such a subset is not “well-posed.” I refer to this general deficiency of arbitrary acausal subsets as the **permeability problem**. As explained below, the discrete causal version of the permeability problem is of great significance in discrete causal dynamics.

A relativistic spacetime manifold is called **globally hyperbolic** if it contains a Cauchy surface. The spacetime W illustrated in the left-hand diagram in Figure 2.7.5 is *not* globally hyperbolic; no acausal subset of W can intersect all inextensible causal curves in W , due to the presence of the “hole” in W . It may be easy to construct acausal subsets intersecting *particular* families of causal curves; for example, the events represented by the four nodes in the figure form a finite acausal subset of W intersecting the four curves γ_1 to γ_4 . However, if one attempts to extend such a subset, one is forced to make choices; for example, the extended subset may intersect certain curves “below the hole” *or* other curves “above the hole,” but not both. The spacetime X illustrated in the right-hand diagram of the figure is globally hyperbolic by definition, since σ is a Cauchy surface. The “global” part of the term “globally hyperbolic” may seem inappropriate, since a particular Cauchy surface represents just one “slice” of spacetime. However, the existence of *even one Cauchy surface* implies much more in the relativistic setting. In particular, a globally hyperbolic relativistic spacetime manifold may be *foliated* by Cauchy surfaces. Informally, this means that the entire spacetime may be viewed as a “stack of Cauchy surfaces;” the individual surfaces are called the *leaves* of the foliation. Each leaf represents a “moment in time” in a particular frame of reference. More precisely, a suitable choice of transition functions defining X as a manifold separates the local temporal variable from the corresponding spatial variables, and the resulting spatial “plaques” patch together across coordinate charts to form the leaves of the foliation. From a dynamical perspective, all the information flowing from past to future may be sampled at any given leaf. This, together with an appropriate dynamical law, enables a “global solution” for the physical behavior modeled by this dynamics; i.e., it allows one to predict “what happens anywhere in spacetime.”

Permeability issues involving acausal subsets. Since discrete causal analogues of Cauchy surfaces are objects of central interest in discrete causal dynamics, it is important to clear up a potential source of confusion regarding these surfaces before proceeding. In comparing the acausal subsets ρ and σ in the right-hand diagram in Figure 2.7.5, it is natural to notice that the “gaps” in ρ may be “plugged” so as to convert ρ into a Cauchy surface, i.e., so as to render it impermeable. From this viewpoint, it is tempting to think of the distinction between an arbitrary acausal subset and a Cauchy surface as one of *completeness*, i.e., to think that a Cauchy surface is roughly the same thing as a “complete” or “maximal” acausal subset. However, this identification is obviously invalid, since only a very restricted class of relativistic spacetime manifolds possess Cauchy surfaces at all. The left-hand diagram in Figure 2.7.6 illustrates, at an informal level, how a maximal acausal subset of a non-globally hyperbolic spacetime may fail to qualify as a Cauchy surface. The acausal subset illustrated here is the union $\rho \cup \rho'$ of two smaller acausal subsets ρ and ρ' of the spacetime manifold W first illustrated in Figure 2.7.5. Here, ρ is represented by the thick black curve “below the hole” in W , which has a “missing point” at x_1 , while ρ' is represented by the shorter thick black curve “above the hole,” which includes its left endpoint x_2 , but not its right endpoint x_3 . The latter subset ρ' is “shielded from ρ by the shadow of the hole,” in the sense that the “hole” prevents causal curves passing through ρ from reaching ρ' . The “shadow of the hole” is represented by the dark gray region.

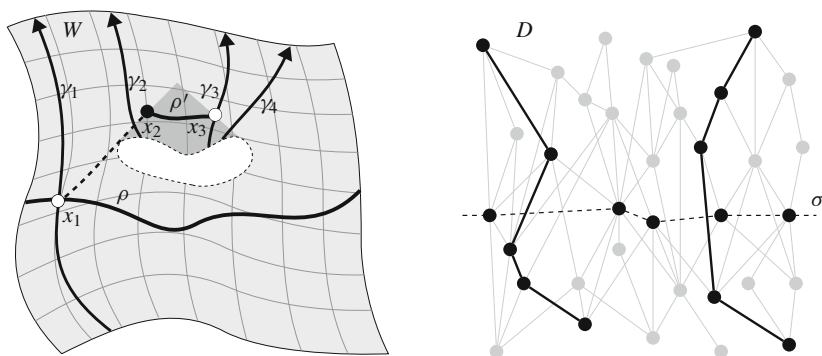


Figure 2.7.6. Permeable maximal acausal subset in relativistic spacetime; a typical maximal antichain in a directed set is highly permeable.

Note that the extension “back in time” of the boundary of this “shadow” intersects the “missing point” x_1 ; i.e., there is a null curve from x_1 to x_2 , represented by the dashed curve in the figure. Hence, the “gap” in $\rho \cup \rho'$ at x_1 cannot be “plugged by adding x_1 ,” the resulting subset $\rho \cup \{x_1\} \cup \rho'$ of W would not be acausal. If the “hole” were absent, then $\rho \cup \{x_1\}$ would be a Cauchy surface. As it is, however, the curves γ_1 , γ_2 , γ_3 , and γ_4 are examples of inextensible causal curves that do not intersect the maximal acausal subset $\rho \cup \rho'$. The curve γ_1 “permeates” $\rho \cup \rho'$ in an obvious way,

since it “goes through the missing point x_1 ,” this curve behaves much like the curves permeating the acausal subset labeled ρ in Figure 2.7.5. The other curves, γ_2 , γ_3 , and γ_4 , “avoid” $\rho \cup \rho'$ in subtler ways, whose details are unimportant at present. The purpose of the illustration is merely to demonstrate how a maximal acausal subset of a relativistic spacetime manifold may fail to be a Cauchy surface.

In discrete causal theory, the permeability problem is a priori “much worse” than in relativity. It occurs more naturally, and is less-strongly tied to global structure. In particular, after defining appropriate discrete causal analogues of Cauchy surfaces in a directed set, one finds that the existence of one such “Cauchy surface” does *not* guarantee that the set may be “foliated by such surfaces.” Further, “most” maximal acausal subsets, called *maximal antichains* in this context, are not even close to being “Cauchy surfaces.” Indeed, they are “riddled with permeations,” rather than merely “missing a few points.” These details are worth remembering when consulting the literature; for example, Bleybel and Zaiour [BZ15] prove a “foliation theorem” for causal sets in a recent paper; however, the “leaves” of the “foliations” involved are *not* close analogues of Cauchy surfaces, since they are generally permeable. Similarly, the *foliation scheme* S_{FOL} introduced in Chapter 7 describes a special type of “generational growth” via generally permeable antichains.

The right-hand diagram in Figure 2.7.6 illustrates a permeable maximal antichain σ in a directed set D , represented by the roughly-horizontal row of nodes connected by dashed lines. These dashed lines are merely a visual aid, included to suggest the “surface-like” characteristics of σ ; they are not part of the actual structure of D . At a mathematical level, the statement that σ is an antichain means that no pair of its elements, distinct or otherwise, is connected by a chain, i.e., a sequence of relations. Chains, which are natural discrete causal analogues of causal curves,²⁹ are studied formally in Chapter 3. The antichain σ is maximal because every other element of D is connected to at least one element of σ by a chain³⁰; hence, it is impossible to add any more elements to σ while preserving its properties as an antichain. The remaining black nodes in the diagram, and the edges connecting them, represent two chains in D permeating σ . The “flow of information” represented by these chains is “invisible to σ ,” and this means that the corresponding initial value problem is not well-posed. The resulting complications for the dynamics of directed sets have already been recognized in the literature, particularly in the special case of causal sets [MRS06]. As described in Chapter 5, passage to relation space provides a pleasing solution to this problem. At least in the acyclic case, the relation space $\mathcal{R}(D)$ over a directed set D is roughly analogous to a “globally hyperbolic spacetime” hidden behind the seemingly intractable structure of D itself. This enables associated dynamical problems to be treated in a much more satisfactory manner.

²⁹It would be more precise to say “analogues of causal *paths*,” see Section 5.9 for details.

³⁰Johnny Feng pointed out to me the a posteriori obvious fact that maximality of an antichain in a multidirected set containing cycles *cannot* be characterized by the condition that “every other element is connected to it by a chain.” For example, according to Definition 3.7.1, the empty subset of a cyclic multidirected set is a maximal antichain, since no element of such a set can belong to an antichain.

2.8 Metric Recovery

Historical context. The previous five sections of this chapter provide enough background information on relativistic spacetime structure and causality conditions to enable the statement of an appropriate version of metric recovery. No formal proof is included, since the details are mostly tangential to the discussion, and may be easily found in the literature. Historically, the theorems of Hawking [HA76] and Malament [MA77], published in the late 1970s, supplied the “critical pieces of the puzzle,” in the sense that they led almost immediately to the birth of discrete causal theory as a serious approach to fundamental spacetime structure. These results have since been amplified in various ways. For example, Hawking and Malament consider only the four-dimensional case, since relativistic spacetime is modeled via four-dimensional manifolds, but the same basic results have since been shown to apply to any dimension at least three. Further, Hawking and Malament assume knowledge of the causal structure of the entire manifold under consideration in the hypotheses of their theorems, but subsequent results have demonstrated that knowledge of the causal structure of a *countable dense subset* suffices. While newer results of this nature are important for establishing the precise details of how causal structure relates to geometry, the original metric recovery theorems are themselves sufficient to motivate the classical causal metric hypothesis (CCMH) in the discrete context.

Statement of the theorem. For the sake of simplicity, I state here a version of metric recovery that may be extracted from Malament’s paper [MA77] alone, although Malament does make use of Hawking’s theorem. The result here is expressed in a somewhat different way than Malament’s main theorem; in particular, it includes two statements, one for enhanced causal isomorphisms of arbitrary relativistic spacetime manifolds, and one for causal isomorphisms of past and future distinguishing relativistic spacetime manifolds. As stated at the end of Section 2.6, Malament works mostly in terms of *timelike* curves, which means that the properties of causal isomorphisms and enhanced causal isomorphisms are actually somewhat stronger than necessary to prove the theorem.

Theorem 2.8.1. Metric recovery. *Let X and X' be smooth four-dimensional real manifolds without boundary, and let g and g' be smooth pseudo-Riemannian metrics of Lorentz signature on X and X' , respectively.*

1. *If $f : X \rightarrow X'$ is an enhanced causal isomorphism, then f is a smooth conformal isometry.*
2. *If X and X' are past and future distinguishing, and $f : X \rightarrow X'$ is a causal isomorphism, then f is a smooth conformal isometry.*

In particular, in either case, knowledge of g enables recovery of g' up to conformal equivalence, and vice versa.

Sketch of Proof. Let X and X' be as described in the statement of the theorem. Informally, the proof involves combining the consequences of the following two statements:

Hawking: “*Topological structure determines conformal structure.*”

Malament: “*Causal structure determines topological structure.*”

Here, I outline the proof as it appears in Malament’s paper [MA77]. Due to Malament’s choice to work mostly in terms of *timelike* curves, certain aspects of the proof are more detailed than necessary to establish Theorem 2.8.1 as I have stated it. However, the additional detail is not too cumbersome, and may help the interested reader to follow the proof in its original context.

The first step in the proof is simply to note Hawking’s result that if $f : X \rightarrow X'$ is a homeomorphism with respect to the manifold topologies on X and X' , and if both f and f^{-1} preserve future-directed continuous null geodesics, then f is a smooth conformal isometry.³¹ Malament refers to Hawking, King, and McCarthy for the proof, while noting that the theorem is described there in a slightly different way.³² The essential argument is described as an “unpublished result of Hawking,” which justifies Malament’s attribution of the theorem to Hawking specifically.

The second step is to extend Hawking’s theorem by means of an easy lemma³³ demonstrating that if $f : X \rightarrow X'$ is a homeomorphism with respect to the manifold topologies on X and X' , and if f and f^{-1} preserve future-directed continuous timelike curves, then f and f^{-1} also preserve future-directed continuous null geodesics.

The third step is to prove that if $f : X \rightarrow X'$ is a bijection, and if f and f^{-1} preserve future-directed continuous timelike curves, then f is a homeomorphism with respect to the manifold topologies on X and X' , and hence, by Hawking’s theorem, a conformal isometry. This is the lengthiest part of the proof; it comprises the entire fifth section of Malament’s paper. This is more than enough to establish the first statement in Theorem 2.8.1. Indeed, if f is an enhanced causal isomorphism, then by Definition 2.6.3, f is a bijection, and f and f^{-1} preserve future-directed continuous *causal* curves, which include future-directed continuous *timelike* curves.

The fourth and final step is to establish that if X and X' are past and future distinguishing, and if $f : X \rightarrow X'$ is a bijection such that f and f^{-1} preserve relativistic *chronological* relations, then f and f^{-1} preserve future-directed continuous timelike curves. Hence, by the previous steps in the proof, f is a conformal isometry. This is more than enough to establish the second statement in Theorem 2.8.1. Indeed, if f is a causal isomorphism, then by Definition 2.6.3, f is a bijection such that f and f^{-1} preserve relativistic *causal* relations, and hence, relativistic chronological relations. □

Topological details. A few technical details are worth clarifying before resuming a more qualitative examination of how metric recovery motivates the classical causal metric hypothesis (CCMH). First, I explain some topological details relating to the proof of Theorem 2.8.1. The reason for emphasizing that the maps involved in the proof are *homeomorphisms with respect to the manifold topologies* on X and X' is

³¹ See Malament [MA77], p. 1400.

³² See Hawking, King, and McCarthy [HA76], p. 174.

³³ See Malament [MA77], Lemma 1, p. 1400.

to avoid potential confusion involving two other topologies that play a role in the papers of Hawking, King, and McCarthy [HA76], and Malament [MA77]; namely, the *Alexandrov topology* and the *path topology*. It is convenient to give here a brief account of these topologies, thereby completing the discussion of the five types of structure on relativistic spacetime, listed in “reverse order of detail” in Figure 2.3.1. Besides assisting the reader in deciphering the literature on metric recovery, this information serves the additional purpose of preparing the ground for later examination of topologies and local properties for directed sets and multidirected sets.

A topology defines which subsets of a set X are “open,” and by complementation, which subsets are “closed.” Most of the details regarding topologies are postponed until Chapter 4. In the present section, I take the informal viewpoint that a topology on X is a collection of subsets of X , called *open sets*, satisfying certain properties abstracted from the properties of intervals on the real line \mathbb{R} . Topologies are particularly useful for describing “local properties of spaces.” Generally, a property is considered to be local near a point x in X if it may be detected by examining any open set containing x , usually called an *open neighborhood* of x . The heuristic that open sets in a topology are analogues of open intervals in \mathbb{R} plays a role both in relativity and in discrete causal theory. In the relativistic case, the Alexandrov topology is a type of *order topology*, or *interval topology*, with respect to the partial order on a relativistic spacetime manifold X satisfying the causal condition. The path topology on X , meanwhile, is defined in terms of *maps* from open intervals in \mathbb{R} into X . In the discrete causal context, however, the naïve idea that “intervals measure local properties,” abstracted from the relativistic setting, leads to serious conceptual and technical issues, as explained in Chapter 4.

The manifold topology on a relativistic spacetime manifold X is inherited from the “usual topology” on \mathbb{R}^4 , via the coordinate charts defining X as a real manifold. This latter topology is the *metric topology* for the usual Euclidean metric on \mathbb{R}^4 . It is defined by taking a subset U of \mathbb{R}^4 to be open if and only if for every element $x \in U$, there exists a positive number ε , such that every element of \mathbb{R}^4 within distance ε of x with respect to the Euclidean metric is also in U . The resulting structure is then transported to X via its coordinate charts in the obvious way. The dark gray “circular” regions illustrated in Figure 2.8.2 represent typical open sets in the manifold topology on a relativistic spacetime manifold X . For \mathbb{R} itself, which may be viewed as one-dimensional Euclidean space \mathbb{R}^1 , the metric topology coincides with the order topology, and hence involves intervals. However, there is no obvious, unique, physically significant partial order on a higher-dimensional Euclidean space, so it is not surprising that the manifold topology on X fails, in certain important ways, to mesh naturally with the physical attributes of X arising from its pseudo-Riemannian structure.

A topology is called “coarse” if it has “few open sets,” and is called “fine” if it has “many open sets.” The **Alexandrov topology** on a relativistic spacetime manifold X is defined to be the *coarsest topology such that the chronological past $I^-(x)$ and the chronological future $I^+(x)$ of each event x in X are open sets*. In particular, every open set in the Alexandrov topology is automatically open in the manifold topology. The “basic open sets” in the Alexandrov topology are the “diamond-shaped” subsets

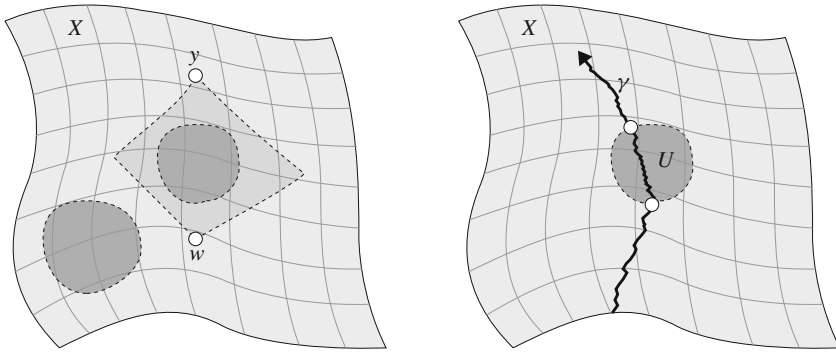


Figure 2.8.2. Manifold topology and Alexandrov topology; defining the path topology.

$I^+(w) \cap I^-(y)$; one of these sets is illustrated in the left-hand diagram in Figure 2.8.2. By the definition of the chronological relation \prec_{GR} on X , these basic open sets are *intervals* of the form $\langle\langle w, y \rangle\rangle := \{x \mid w \prec_{\text{GR}} x \prec_{\text{GR}} y\}$. From a physical standpoint, the basic open set $\langle\langle w, y \rangle\rangle$ consists of all events that may be reached from the event w by the motion of a massive body, and may also reach the event y by the motion of a massive body. This description demonstrates the obvious physical significance of the Alexandrov topology in the relativistic setting.

A topology on a set X *induces* topologies on each of its subsets S , by defining a subset V of S to be open if and only if $V = U \cap S$ for some open subset U of X . The **path topology** on a relativistic spacetime manifold X is defined to be *the finest topology that induces, on the images of all continuous timelike curves in X , the same topology induced by the manifold topology*. In particular, every open set in the manifold topology is automatically open in the path topology. The right-hand diagram in Figure 2.8.2 illustrates how an open subset U of X in the manifold topology, represented by the dark “circular” region, defines an open subset $V = U \cap \gamma$ of the image of a continuous timelike curve γ in X , represented by the “part of γ inside U ,” which has its “endpoints missing.” The physical significance of the path topology is slightly less direct than that of the Alexandrov topology; continuous timelike curves are obviously significant in their own right, but the finest choice of “open sets” in X that reproduces the usual manifold-induced topology on the images of such curves is a bit of a nuisance to describe. The real advantage of the path topology is that its homeomorphisms are precisely the smooth conformal isometries of X . To paraphrase Malament,³⁴ the path topology “*simultaneously encodes information about the manifold structure, the smooth structure, and the conformal structure of X .*”

Modern improvements on metric recovery. A second technical point, already mentioned in passing above, is that the strength and scope of metric recovery results has been significantly improved since the original metric recovery theorems of the

³⁴Malament [MA77], p. 1399.

late 1970s. For example, Luca Bombelli and David Meyer’s 1989 paper *The origin of Lorentzian geometry* [BO89], and Keye Martin and Prakash Panangaden’s 2006 paper *A Domain of Spacetime Intervals in General Relativity* [MP06], both demonstrate metric recovery results for which much weaker hypotheses, involving only the causal structure of a *countable dense subset* of a relativistic spacetime manifold, suffice. Metric recovery has also been extended to pseudo-Riemannian manifolds of Lorentz signature in any dimension at least three; for a discussion of this, see the recent paper of Parrikar and Surya [PS11]. Since many proposed theories attempting to improve upon general relativity, especially string theory and M -theory, make use of higher-dimensional manifolds, these results are of more than academic interest. The fact that analogous results fail to hold in dimension two is also interesting; for example, because of simulations suggesting *dimension reduction* in certain fields of quantum gravity, and because of the prominence of conformal field theories on Riemann surfaces in other areas of theoretical physics.

Motivation for the causal metric hypothesis. The simplest way to express the meaning of metric recovery in the context of general relativity is to say that *causal structure determines metric structure up to scale*. As discussed in Section 2.3, this statement approaches, but does not quite reach, the elegant and tempting conclusion represented by the classical causal metric hypothesis (CCMH); namely, that *metric structure is merely an approximate way of describing causal structure*. The qualifier “up to scale” obstructs such a conclusion in the relativistic case, but the idea is sufficiently compelling that it is natural to ask if one can somehow justify it by shifting attention to “causal structures possessing a natural scale.” As mentioned in Section 1.3, and again in Section 2.6, the founders of causal set theory perceived the most obvious way to accomplish this, almost immediately after the original metric recovery theorems were established: to work with discrete models, and to assign volume to subsets by counting their elements. This is the strategy that Sorkin later encapsulated in his phrase, “*order plus number equals geometry*,” where “order” stands for causal structure, and “number” stands for the counting procedure. Perhaps the most straightforward way to realize this idea is by means of causal sets constructed via “sprinklings” into pseudo-Riemannian manifolds, such as Minkowski spacetime \mathbb{R}^{3+1} . Such causal sets are discussed briefly in Section 3.2, and more thoroughly in Section 4.5.

Options for realizing the hypothesis. However, the causal set approach is only one of many possible ways to realize the classical causal metric hypothesis, even if one restricts attention to the discrete context. “Order” is a very restrictive, and likely inadequate, proxy for causal structure, even in the relativistic case. In particular, the relativistic causal relation $<_{\text{GR}}$ defines a partial order only for relativistic spacetime manifolds satisfying the causal condition. “Number,” meanwhile, is a very specific, and possibly oversimplified, proxy for scale. A general hazard to be avoided when converting a compelling conceptual motif; in this case, the classical causal metric hypothesis, into a specific technical approach; for example, causal set theory, is the risk of ignoring equally viable approaches that may ultimately reach further. Hence, it is crucial not to narrow down the possibilities prematurely. In particular, even if

one restricts attention to spacetimes satisfying the causal condition, the transition to the discrete context is nontrivial, and automatically transferring over all the axioms of partially ordered sets is unjustified. Similarly, it is by no means obvious that each element in a discrete directed set should contribute equally, or even approximately equally, to the “volume” of the set. After all, one of the most obvious ways in which a typical such a set differs from a manifold is that it is *locally irregular*; i.e., its local structure is generally *not* the same near each element.

The only truly essential feature of causal structure, beginning from first principles, is the directed relationship between cause and effect. A partial order takes this *local* building block of structure, and adds additional *nonlocal* properties that are not necessarily appropriate; for example, transitivity (TR). Similarly, the only truly essential requirement regarding the introduction of scale data in the context of the classical causal metric hypothesis is that this data should arise *naturally* from the causal structure itself. Following these lines of thought, it is useful to juxtapose Sorkin’s version of the classical causal metric hypothesis with a more general statement that, while less succinct, avoids the risk of placing the subsequent technical development in a structural straitjacket:

Sorkin: “*Order plus number equals geometry.*”

Generalization: “*Directed structure plus natural scale equals geometry.*”

Figure 2.8.3 illustrates three different ways of assigning “volume data” to a discrete directed set. The left-hand diagram illustrates Sorkin’s original prescription, taken literally; every element is assigned exactly the same volume. The middle diagram illustrates the incorporation of “statistical fluctuations” in the computation of volume. Such fluctuations are invoked in the causal set literature for technical reasons; in particular, to avoid systematic violations of Lorentz invariance for causal sets constructed via global “sprinkling” into Minkowski spacetime \mathbb{R}^{3+1} . The right-hand diagram illustrates a much different method of assigning volume, in which not only the number of elements, but also the *local causal structure*, plays a role. In this particular case, the “volume” of each element is determined by its *valence*, i.e., by the number of relations for which it is the initial or terminal element.³⁵ In this context, the valence field illustrated in Figure 1.7.2 of Chapter 1 serves as a “volume field,” i.e., a “discrete conformal factor.”

From the *relation space viewpoint*, introduced briefly in Section 1.5, and developed in detail in Chapter 5, the latter method of assigning scale data is perhaps the most attractive of the three rather naïve methods illustrated in the figure. This is because relations, rather than elements, are considered to be fundamental in this context, and the valence field essentially “counts relations.” However, discrete directed sets are rich in combinatorial structure, and there are many other possible methods of deriving scale data from this structure. In a way, this is a disadvantage, because

³⁵ As explained in Section 4.3, *reflexive relations* $x \prec x$ are counted twice in enumerating $v(x)$, because such relations both begin and terminate at x . Of course, such relations do not occur in *acyclic* directed sets.

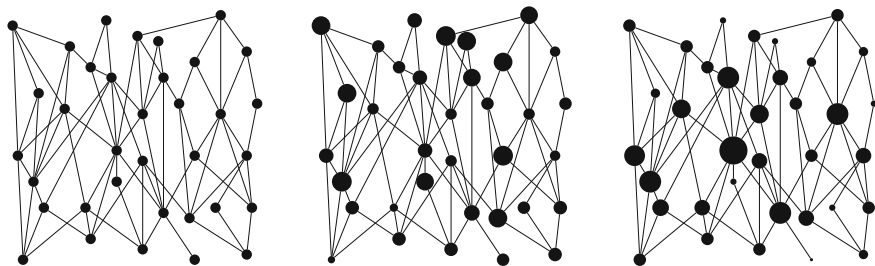


Figure 2.8.3. Alternative methods of assigning volume to a discrete causal structure: constant; incorporating statistical fluctuations; valence-based.

it introduces a risk that the selection of one particular method among these will be unconvincing. It would be preferable if one of the simplest methods, involving a straightforward counting of elements or relations, could be proven to be the “right one.” However, there are other quantities just as basic as scale from a geometric viewpoint, such as dimension, whose emergence from discrete causal structure almost certainly does not arise from a method as direct as a simple counting procedure. Perhaps the most reasonable approach to this situation, while discrete causal theory remains in a relatively early stage of development, is to focus on the simplest models, while keeping in mind the possibility that subtler considerations may ultimately be incorporated as “corrections” in more mature versions of the theory. This strategy is supported by the plausible reflection that a variety of different methods of assigning scale data may very well “converge” within just a few orders of magnitude of the fundamental scale. In particular, the “variable volume” method illustrated in the right-hand diagram in Figure 2.8.3 could, in many cases, be replaced by a simpler causal set-type method, using the “average valence” of the elements, leading to only negligible differences in the resulting computations above the very smallest scales. In fact, if the typical valences of elements in a physically realistic directed set turn out to be very large, then the differences between the volumes assigned to elements under a valence-based approach to volume could be negligibly small in proportion even at the fundamental scale itself. Finally, Bombelli, Henson, and Sorkin [BHS09] suggest the possibility of adding “distance information” to the relations of a causal set, but conjecture that causal sets represent “*in some sense, the minimal [systematically] Lorentz invariant discrete [structures] from which [Minkowski spacetime] can be reconstructed at macroscopic scales.*”

2.9 Order Good, Continuum Bad

Criticizing continuum-based theories. The foregoing sections of this chapter provide preliminary evidence suggesting that discrete causal theory is a reasonable candidate to serve as an alternative structural paradigm for modeling classical space-

time. However, it is worthwhile to consider the basic question of *why such alternatives are needed in the first place*; in particular, why real manifolds are ultimately not ideal for this role, despite their distinguished history in theoretical physics. In this section, I present a perspective on this subject that might be considered “underrepresented in the literature;” namely, that *the prominence of continuum-based theories is partly just a historical accident*, resulting from broad shortcomings in human knowledge, and human computational capabilities, throughout the early development of modern science. From this point of view, there is nothing natural or inevitable about the role of the real numbers to recommend them as the default building block of basic physical structure to any sufficiently advanced scientific community. In particular, certain well-known obstructions to progress in standard continuum-based theories, including divergence issues and problems with renormalizability, are regarded, from this perspective, less as “great problems to be solved,” than as evidence that the wrong questions are being asked, based on the wrong presuppositions. Discrete causal theory aims, as far as possible, to completely circumvent such obstructions.

A nonspecific hypothesis of fundamental discreteness no longer seems *avant-garde* in its own right, since it is now widely expected that “the spacetime continuum breaks down at the fundamental scale” in the context of quantum gravity. Indeed, a large number of different *non*-continuum-based approaches to fundamental physics, and more specifically, discrete approaches, may be found in the literature.³⁶ However, even a cursory examination of leading publications suggests that the vast majority of mainstream modern theoretical physicists still devote their efforts almost exclusively to theories that use the real numbers as a structural “starting point” in one form or another. In particular, string theory, which remains, by a huge margin, the dominant approach, both in terms of its number of researchers and its quantity of resources, is entirely continuum-based. Its main competitor, loop quantum gravity, “arrives” at a form of fundamental discreteness in a circuitous manner, via a novel quantization procedure, beginning in the continuum-based context of general relativity.³⁷ Hence, even though it is generally believed that “quantum spacetime is not a real manifold,” the actual practice of modern theoretical physics mostly fails to reflect this belief.

Many of the specific technical problems arising in continuum-based theories are completely irrelevant in the discrete causal context. For example, *renormalization* is a standard device in quantum field theory for “curing” certain divergence issues that arise, ultimately, from properties of the real numbers. This approach succeeds only in special cases; theories for which it fails are called *nonrenormalizable*. One of the major historical obstacles in formulating successful theories of quantum gravity has been the fact that “standard” approaches to “quantizing general relativity” lead to nonrenormalizable theories. In discrete causal theory, however, the original sources of the divergences eliminated by renormalization are generally absent, and

³⁶An interesting list of such approaches appears in the recent paper on *tensor networks* by Chen, Sasakura, and Sato [CSS16]. The majority of these approaches are less relevant to this book than causal set theory, since they rely to a large degree on auxiliary, “non-causal” structure.

³⁷Of course, as discussed in Section 8.8, much of this reliance on the continuum may be stripped away *a posteriori*.

this renders the whole subject of such devices a priori irrelevant, at least in its original context.³⁸ Of course, objects such as renormalization groups, or analogues thereof, may reappear in interesting mathematical roles, possibly intersecting with discrete causal theory. As originally conceived, however, such methods may be regarded as “chemotherapy for continuum-based theories,” i.e., as tortuous cures for problems that one would prefer to avoid entirely.

Similarly, the deeply-entrenched and tiresome controversy over the *string theory landscape*, and its implications regarding the *anthropic principle*, arises ultimately from the properties of certain families of manifolds, including the iconic *Calabi–Yau manifolds*, which are likely of greater mathematical than physical interest. These manifolds have been “imported” into string theory from algebraic geometry, in order to “cure” the inconvenient fact that string theory requires the wrong dimension for spacetime, as indicated by all available observational evidence. A serious description of such problems would contribute nothing to the subject of this book, so I choose instead to focus on more basic and foundational objections to the entire corpus of continuum-based theories, in particular, objections arising from generic structural properties of real manifolds that are almost certainly physically irrelevant. Many of the technical struggles of modern theoretical physics, including those mentioned above, may be regarded as mere symptoms of these deeper problems.

The real numbers \mathbb{R} ; order and completeness. Every physicist is familiar with the real number system \mathbb{R} , which supplies the structural scaffolding for the continuum-based theories that dominate conventional modern physics. \mathbb{R} is a *linear continuum*, in a sense made precise below; for the moment, it suffices to remark that the word “linear” refers to a purely order-theoretic property of \mathbb{R} , while the word “continuum” refers to a “completeness property,” which is also *essentially* order-theoretic, but which may be generalized to apply to non-ordered sets. In this book, the term *continuum-based theory* refers to a physical theory described in terms of real manifolds. Such manifolds inherit a completeness property from \mathbb{R} , but generally possess no natural order. Hence, in passing from \mathbb{R} itself to manifolds defined over \mathbb{R} , one loses most of the original order-theoretic structure. To construct continuum-based theories that yield even an approximate description of nature, one must “add back in” order-theoretic structure artificially, by means of a metric. In the discrete causal context, order-theoretic properties are essential, because the directed sets used to model discrete causal classical histories derive their local structure from individual ordered relationships between pairs of elements. Hence, even though order theory is not quite general enough to model classical causal structure in a global sense, it remains part of the conceptual core of the theory.

Despite the universal familiarity of the real numbers, it is instructive to re-examine \mathbb{R} in detail as a *mathematical object*, but with a view toward physical applications. This enables a better understanding of some of the basic objections to continuum-based theories of fundamental physics. These objections strongly suggest a need for alternative approaches, of which discrete causal theory is an obvious candidate. It

³⁸See, however, the discussion in Section 4.5 regarding “sprinkled” causal sets, which may exhibit locally infinite behavior.

turns out that most of the order-theoretic structure of \mathbb{R} may be regarded as “good” in this context, with totally ordered and partially ordered sets retaining a prominent role in discrete causal theory. However, the completeness property of \mathbb{R} , which is the property that “makes it a continuum,” leads to deep trouble. In fact, even the weaker *interpolative property* of \mathbb{R} , which says that “one can always find a real number strictly between any two distinct real numbers,” is problematic. In particular, even the “incomplete” field of rational numbers \mathbb{Q} is already “bad” from the discrete causal perspective, since it shares the interpolative property with \mathbb{R} .

\mathbb{R} in terms of universal properties. Standard definitions of \mathbb{R} , though given in almost every university analysis course, appear surprisingly obtuse when approached from first principles. \mathbb{R} may be defined, in terms of universal properties, as *the unique isomorphism class of Archimedean complete totally ordered fields*. To review a bit of algebra, a field is an “optimally behaved number system,” in the sense that it possesses additive and multiplicative operations that satisfy familiar properties, and that “cooperate” with each other in familiar ways. To be precise, a **field** is a set \mathbb{F} , together with two operations $+$ and \times , called *addition* and *multiplication*, respectively, such that $(\mathbb{F}, +)$ is an abelian group with identity 0, $(\mathbb{F} - \{0\}, \times)$ is an abelian group with identity $1 \neq 0$, and multiplication distributes over addition.

Total order. A *total order*, or synonymously, a *linear order*, on a field \mathbb{F} , is a *transitive*, *antisymmetric*, *total* binary relation \leq on \mathbb{F} . Letting x , y , and z be elements of \mathbb{F} , “transitive” means that if $x \leq y$ and $y \leq z$, then $x \leq z$, “antisymmetric” means that if $x \leq y$ and $y \leq x$, then $x = y$, and “total” means that either $x \leq y$, or $y \leq x$ for every choice of x and y , including $x = y$. In particular, \leq is *reflexive*; i.e., $x \leq x$ for every $x \in \mathbb{F}$. Omitting the “total” property, but retaining reflexivity, yields a *partial order*. The set of real numbers \mathbb{R} is a totally ordered field under the familiar “less than or equal to” relation \leq . Given a total order \leq on \mathbb{F} , one may define a unique irreflexive binary operation $<$ on \mathbb{F} , sometimes called a *strict total order*, by setting $x < y$ if and only if $x \leq y$ and $x \neq y$. The familiar “less than” relation $<$ on \mathbb{R} is the strict total order corresponding to \leq . Conversely, given a strict total order $<$ on \mathbb{F} , one may define a unique non-strict total order \leq on \mathbb{F} in the obvious way, by setting $x \leq y$ if and only if $x < y$ or $x = y$. It is convenient here to dispose of some nuisances of terminology. First, strict orders such as $<$ are more useful in this book than nonstrict orders such as \leq , even though the latter are often more popular in mathematical settings. Second, partial orders play a larger role in this book than total orders. Hence, most encounters with “order theory” in this book actually involve *strict partial orders*, whether or not the words “strict partial” appear explicitly. By contrast, the discussion of ordered fields in the present section follows the usual mathematical conventions; in particular, the “order” on \mathbb{R} is taken to be the usual non-strict total order \leq .

It is important to note that the total order \leq on \mathbb{R} is a natural aspect of the structure of \mathbb{R} , not an arbitrary auxiliary structure added a posteriori. Following the order refinement principle (ORP), discussed in Section 3.8, any set, and hence any field, may be endowed with a total order. For example, the field of complex numbers \mathbb{C} , whose elements are of the form $a + bi$, where a and b are real numbers and

i means the imaginary unit $\sqrt{-1}$, may be endowed with the total order borrowed from the familiar *lexicographic order* on \mathbb{R}^2 , i.e., the order \leq defined by setting $a + bi \leq c + di$ if and only if $a < c$, or $a = c$ and $b \leq d$, under the usual order on \mathbb{R} . The distinction between this sort of ad hoc total order imposed on a field, and a natural total order, such as the usual order \leq on \mathbb{R} , is that the latter “respects the field structure.” For example, given positive elements w, x, y , and z of \mathbb{R} , it is true that if $w \leq x$ and $y \leq z$, then $wy \leq xz$. The analogous property fails to hold for the total order on \mathbb{C} defined above; for example, let $w = y = 1$ and $x = z = 1 + i$; then $wy \not\leq xz$, since the real part of wy is 1 and the real part of xz is 0. Generally, when one speaks of an “ordered” algebraic object, the order is assumed to respect the algebraic structure, unless stated otherwise.

Archimedean property. The order \leq on a totally ordered field \mathbb{F} provides a way of comparing any pair of elements x and y of \mathbb{F} . In particular, any nonzero element x satisfies either $x < 0$ or $0 < x$ under the corresponding strict total order; in the first case, x is called *positive*, and in the second case, x is called *negative*. The Archimedean property says, informally, that “given any pair of positive elements x and y in \mathbb{F} , either element may be re-scaled to become larger than the other.” More precisely, a positive element x is called *infinitesimal* with respect to a positive element y if every natural-number multiple of x is less than y ; the Archimedean property says that \mathbb{F} has no pairs x and y such that x is infinitesimal with respect to y . To readers without much background in abstract algebra, the Archimedean property may seem “obvious,” but there exist familiar and important examples of non-Archimedean total ordered fields. For example, the field $\mathbb{R}(x)$ of rational functions in one variable x , with real coefficients, possesses a natural total order, defined in terms of the leading coefficients of numerator polynomials. However, it possesses infinitesimal elements; for example, $1/x$ is infinitesimal with respect to 1.

Completeness. The remaining property of \mathbb{R} cited above is *completeness*, and it is this property that distinguishes \mathbb{R} as the only continuum, up to isomorphism, among the class of Archimedean totally ordered fields. The set of rational numbers \mathbb{Q} , for example, is an Archimedean totally ordered field, but it is not complete, since it “leaves out” certain “limiting values,” such as the algebraic number $\sqrt{2}$, and the transcendental numbers π and e . Completeness is defined in terms of *Cauchy sequences* in a totally ordered set, which are sequences whose elements “eventually become arbitrarily close to each other,” in a manner familiar from elementary calculus. However, a potential problem of self-reference arises in this context, because Cauchy sequences are usually defined *in terms of* \mathbb{R} , which, of course, is the object of present scrutiny. In more detail, Cauchy sequences in a set S are usually defined in terms of a generalized distance function d , called a “metric,” which quantifies the “closeness” of pairs of elements of S . To be clear, such a “metric” d does *not* represent the same type of structure as a metric g on a smooth real manifold; e.g. a pseudo-Riemannian metric, although a metric of Euclidean signature *induces* a “metric” d in the present sense. Rather, d is a map $S \times S \rightarrow \mathbb{R}$, which is *positive-definite*, *symmetric*, and satisfies the *triangle inequality*. Defining “closeness” with respect to such a “metric” d is circular when discussing the real numbers themselves, since the target of d is \mathbb{R} .

A more general notion is needed in this context, and this is supplied by the totally ordered group structure of $(\mathbb{R}, +)$, which defines what is called a *uniform structure*. Generalized Cauchy sequences may be defined with respect to any uniform structure, and this method may be used to characterize the completeness of \mathbb{R} . In general, a set S endowed with a uniform structure is called *complete* if every Cauchy sequence defined in terms of this structure converges to an element of S . The real numbers are complete with respect to the uniform structure defined by $(\mathbb{R}, +)$.

\mathbb{R} as a linear continuum. Technically, a *linear continuum* L is an *interpolative totally ordered set satisfying the least-upper bound property*. Here, letting x , y , and z be elements of L , and denoting the strict total order on L by $<$, “interpolative” means that for any pair of elements x and z in L with $x < z$, there exists a third element y “between the two,” i.e., such that $x < y < z$. The least upper bound property says, as one would expect, that any subset S of L that is bounded above has a least upper bound in L . More precisely, a subset S of L is bounded above if there exists an element $u \in L - S$ such that $s \leq u$ for every $s \in S$. A least upper bound for S is an element $u_{\text{MIN}} \in L$ that is, first of all, an upper bound of S , and secondly, is less than any other upper bound of S . By antisymmetry, u_{MIN} is necessarily unique if it exists. The rational numbers \mathbb{Q} fail to satisfy the least upper bound property; for example, the subset of rational numbers $\{\frac{1}{1}, \frac{3}{2}, \frac{8}{5}, \frac{13}{8}, \dots\}$, defined in terms of the Fibonacci numbers, is bounded above by the rational number 2, by an easy induction argument, but does not possess a rational least upper bound. Its least upper bound in \mathbb{R} , of course, is the golden ratio $\phi = \frac{1+\sqrt{5}}{2}$.

In the special case of Archimedean totally ordered fields, completeness is enough to guarantee the interpolative property and the least upper bound property defining a linear continuum, but in more general contexts, completeness does not imply continuum structure. For example, the set of integers \mathbb{Z} is complete, since every Cauchy sequence is “eventually constant” at a specific integer n , to which it therefore converges. However, \mathbb{Z} is not a continuum, because it does not satisfy the interpolative property. Even for the rational numbers, there exist different, non-Archimedean, completions, which are not continua; namely, the p -adic fields \mathbb{Q}_p , in which “distance” is quantified in terms of divisibility properties. It is interesting to note that there has recently emerged an entire field of fundamental physics devoted to non-Archimedean versions of quantum theory and related topics, sometimes called *p -adic quantum mechanics*.

Real manifolds as “unordered continua.” As suggested above, the word “continuum,” without the qualifier “linear,” is often used in a general manner to refer to certain spaces that possess an appropriate “completeness property,” but which generally do not possess natural order-theoretic structure. For example, a *topological continuum* is defined to be a compact, connected topological space equipped with a “metric,” i.e., a generalized distance function like the “metric” d discussed above. In this context, the necessary completeness property is embodied by the compactness condition, which may be expressed in terms of the convergence properties of generalized sequences. However, in the context of theoretical physics, the word “continuum” is often used as a synonym for “real manifold.”

Real manifolds are generally neither linear continua nor compact connected metric spaces. In particular, essentially the only “ordinary” real manifold possessing the natural structure of a linear continuum is \mathbb{R} itself,³⁹ while an n -dimensional real manifold is, by definition, locally isomorphic to \mathbb{R}^n , which itself possesses no natural order-theoretic structure for $n \geq 2$. However, real manifolds are “locally complete,” in the sense that every point in a real manifold X possesses a neighborhood which “contains all limit points of sequences in the neighborhood.” This is illustrated in Figure 2.9.1, which shows a sequence $\{x_n\}_{n \in \mathbb{N}}$ of points converging to a point x in X . Order theory plays only an indirect role in this notion of completeness; the points of the sequence are ordered, but this order is borrowed from the natural numbers \mathbb{N} , and does not reflect any essential structural aspect of X itself. At a formal level, the sequence $\{x_n\}_{n \in \mathbb{N}}$ may be viewed as a map from \mathbb{N} into X , and this is how it is represented in the figure.⁴⁰ A *path* in a set X , in a generalized order-theoretic sense, is an equivalence class of maps from a “linear directed set” into X , so this sequence may be regarded as representing a “discrete path” in X . However, it is not a “discrete directed path,” i.e., an equivalence class of morphisms of directed sets from \mathbb{N} into X , since X is not assumed to be a directed set in this context. Indeed, the only structure on X taken for granted here is its real manifold structure.

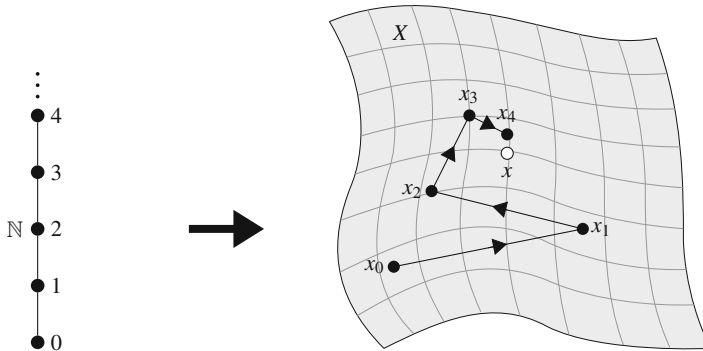


Figure 2.9.1. A sequence $\{x_n\}$ converging to a point x in a real manifold X , viewed as a map $\mathbb{N} \rightarrow X$.

It is both unfortunate and ironic that the ubiquitous use of the word “continuum” to mean “real manifold” in theoretical physics *emphasizes only the completeness property* of the linear continuum \mathbb{R} , the object from which every real manifold derives most of its structure, and *suppresses the role of the other order-theoretic properties*

³⁹The informal qualifier “ordinary” is included to rule out “exotic examples” such as the *long line*.

⁴⁰The directed set illustrated in the left-hand diagram in Figure 2.9.1 is really the *skeleton* $\text{sk}(\mathbb{N})$ of \mathbb{N} , not \mathbb{N} itself, since reducible relations are not included in the figure. Skeletons are introduced in Chapter 3.

of \mathbb{R} . This completeness property, despite its mathematical convenience in the context of calculus and differential equations, is probably the least relevant and most problematic feature of \mathbb{R} from a physical standpoint. From the perspective of the causal metric hypothesis (CMH), it is the *linear order* of \mathbb{R} that stands out as its principal redeeming characteristic, since this order enables subsets of real intervals to parameterize sequences of events.

Constructive view of \mathbb{R} . Of course, mainstream modern physics devotes little attention to the basic properties of \mathbb{R} , whether order-theoretic, topological, analytic, or algebraic. Instead \mathbb{R} is usually treated as a collection of numbers which can serve, at least in principle, as the values of measurements, and which is moreover *large enough and well-enough behaved* to enable convenient methods of mathematical analysis; in particular, calculus and differential equations. Here, the viewpoint represented by the phrase “large enough and well-enough behaved” may be understood in terms of the historical emergence of \mathbb{R} as a “number system,” which encourages the unfortunate impression that “progress” involves a constructive process of adding more and more structure until “enough structure is present to describe nature.” In fact, such a process risks, at each step, the inclusion of mathematically convenient but physically irrelevant structure, which may ultimately lead the researcher, or the entire scientific community, in the wrong direction.⁴¹

The history of the constructive process leading to \mathbb{R} might be told in the following naïve way: the positive integers, which had been used since paleolithic times to count quantities such as the number of bear skins harvested in a given month, required augmentation by a zero element, to facilitate such innovations as the place value system; this led to the natural numbers \mathbb{N} . Similarly, negative integers and fractions were added to the picture to describe such concepts as directions on the number line, and equal partitions of objects; this yielded the integers \mathbb{Z} and the rational numbers \mathbb{Q} . Early geometry, and later algebra, raised awareness of the existence of pairs of idealized quantities not in integer ratios to each other; for example, the Greeks noticed that the diagonal of a square in the Euclidean plane is incommensurable with its edges. This eventually motivated the recognition of irrational numbers such as $\sqrt{2}$, and transcendental numbers, such as π and e . The table in Figure 2.9.2 shows some of the “number systems” appearing in this process, along with their basic algebraic and order-theoretic properties. As indicated by the three bottom rows of the table, the process may be continued “beyond \mathbb{R} .”

⁴¹It is prudent to balance this warning with the *principle of hidden structure (HS)*, introduced in Chapter 3, which emphasizes the utility of *enlarging* the structural picture, if necessary, with new features that *are* relevant! The moral, ultimately, is that one must always keep in mind the physical motivation, or lack thereof, of structural features in physical theories.

symbol	name	algebraic structure	order structure
$\mathbb{N} - \{0\}$	positive integers	semigroup under $+$, monoid under \times	discrete total order
\mathbb{N}	natural numbers	monoid under $+$, monoid under \times	discrete total order
\mathbb{Z}	integers	ring (under $+$, \times)	discrete total order
\mathbb{Q}	rational numbers	field	interpolative total order
\mathbb{R}	real numbers	field	linear continuum
\mathbb{C}	complex numbers	algebraically closed field	no natural order
\mathbb{H}	quaternions	non-commutative division algebra	no natural order
\mathbb{O}	octonions	non-associative division algebra	no natural order

Figure 2.9.2. Comparison of “number systems,” increasing in “size” from *top* to *bottom*.

From this constructive viewpoint, \mathbb{R} may ultimately be defined as *the set of equivalence classes of Cauchy sequences of rational numbers*, under an appropriate equivalence relation. For example, the sequence $\{\frac{1}{1}, \frac{3}{2}, \frac{7}{5}, \frac{17}{12}, \dots\}$ is a representative of the equivalence class identified with $\sqrt{2}$.⁴² Now suppose that there really is a minimum physical length; for example, the Planck length, and suppose that one wishes to “describe the diagonal of a macroscopic square-shaped region in flat space in terms of its edge.” The naïve description involves the irrational number $\sqrt{2}$, but the precise physical description must be rational; for instance, there might be exactly 1.414213562373095048801688724209698 units of length along the diagonal for each unit of length along the edge. This example illustrates one reason why hypotheses involving discrete microstructure might generate little sympathy from a naïve point of view, since it is often a waste of time, practically speaking, to worry about so many decimal places. It might seem that “nothing is lost” by enlarging the set of possible values to admit quantities such as $\sqrt{2}$, even if they may not be precisely physical.

Probably no serious modern theorist actually holds a view so naïve as this, but the example nonetheless provides a reasonably accurate illustration of some of the conceptual pitfalls surrounding the physical role of \mathbb{R} , which *do* seem to exert a profound influence on the way mainstream theoretical physics is done. In particular, the impression that “nothing is lost” by enlarging, or completing, a number system, is egregiously wrong in general. In the context of classical causal structure, the nat-

⁴²The method of *Dedekind cuts* is a different way of constructing \mathbb{R} in terms of subsets of \mathbb{Q} .

ural scale is lost thereby, and with it, metric recovery. In this setting, it is striking to consider the fact that smooth manifolds, by themselves, are *manifestly* inadequate for modeling classical spacetime; even in relativity, one must import a metric as auxiliary structure. Discrete directed sets, by contrast, can *at least* approximate classical spacetime to a high degree of accuracy, even if spacetime is ultimately nondiscrete. In a broader setting, an unfortunate tendency exists to automatically place objections to the *physical* status of \mathbb{R} in the same category as ancient Greek qualms about irrational numbers, or the constructivist arguments of Kronecker, Brouwer, and others, made more than a century ago. In fact, the objections to \mathbb{R} most relevant to the motivations for discrete causal theory have nothing to do with the ontology of number systems, or with mathematical constructivism more generally.

Larger related number systems. Before enumerating some of these specific objections, I briefly outline the remaining content of the table in Figure 2.9.2, including the “larger number systems” \mathbb{C} , \mathbb{H} , and \mathbb{O} . First, I must explain the algebraic terminology appearing in the table. A *semigroup* is a set equipped with an associative binary operation, but generally without an identity or inverses. The prototypical example is the set $\mathbb{N} - \{0\}$ of positive integers⁴³ under addition, since, informally speaking, the “obvious candidates for an additive identity and additive inverses,” namely, the zero element and the negative integers, respectively, are “left out.” A *monoid* is a semigroup with identity; the prototypical examples are the positive integers under multiplication, with identity 1, and the natural numbers \mathbb{N} under addition, with identity 0. A *ring* is “like a field,” in the sense that it possesses “additive” and “multiplicative” operations, but it is more general; in particular, nonzero elements of a ring may not possess multiplicative inverses, multiplication may not be commutative, and so on. The prototypical example of a ring is the set of integers \mathbb{Z} , with the usual multiplication and addition. The construction of \mathbb{Q} from \mathbb{Z} is algebraic in nature; it is an example of what is called *localization*. The terminology arises from the fact that the same construction is used to examine local properties of algebraic schemes in algebraic geometry. The construction of \mathbb{R} from \mathbb{Q} is order-theoretic, as described above.

The complex numbers form an algebraically closed field, which means that the roots of any polynomial with coefficients in \mathbb{C} are also in \mathbb{C} . Hence, the construction of \mathbb{C} from \mathbb{R} is algebraic. The complex numbers play a central role in ordinary quantum theory; in particular, the “state function” ψ appearing in Schrödinger’s equation (1.1.3) takes on complex values, and the phases associated with spacetime paths in Feynman’s path summation approach to quantum theory are also complex-valued. The apparent physical importance of progressively larger number systems, up to and including \mathbb{C} , has provoked natural interest in the even larger systems of *quaternions* \mathbb{H} and *octonions* \mathbb{O} , the latter of which is the “largest normed division algebra over the real numbers.” For example, since passage from \mathbb{R} to \mathbb{C} “produces quantum theory from classical theory,” it is natural to ask if there might exist “hyperquantum

⁴³This book follows the convention that the natural numbers \mathbb{N} include zero; hence, the set of positive integers is given by removing zero from \mathbb{N} .

theories,” based on \mathbb{H} and \mathbb{O} .⁴⁴ Examination of these larger number systems has borne only modest fruit in physical settings; both \mathbb{H} and \mathbb{O} play a limited role in quantum information theory, and \mathbb{H} makes appearances in spin dynamics and a few other contexts. However, these contributions remain miniscule compared to those of \mathbb{R} and \mathbb{C} .

The table of “number systems” in Figure 2.9.2 may be placed into a larger context, in which “simple objects,” such as \mathbb{N} , and even “simpler objects,” such as finite sets, may be generalized and/or extended in a variety of different ways. For example, finite sets may be converted into “number systems,” by endowing them with group or ring structure. Another example is given by rings of *adeles*, which generalize the completion \mathbb{Q}_p of \mathbb{Q} . Examples more directly relevant to discrete causal theory are *ordinal numbers* and *cardinal numbers*, which generalize naïve “counting” of finite sets in different ways. It is sometimes useful to regard directed sets as generalizations of ordinal numbers; for example, this viewpoint is useful in the theory of *relative directed sets over a fixed base*, introduced in Section 4.6. From this perspective, Sorkin’s phrase “*order plus number equals geometry*” “almost” says that “*ordinal plus cardinal equals geometry*,” since the partial orders involved in causal set theory generalize the total orders defining ordinal numbers, while the “quantity,” or “measure,” or “scale,” represented by Sorkin’s use of the word “number,” refines the “pure size” meaning associated with cardinal numbers.

Quantum-theoretic significance of \mathbb{C} . A common heuristic is to associate the real numbers \mathbb{R} with “classical physics,” and the complex numbers \mathbb{C} with “quantum physics.” For example, the reader may recall that the use of *real* probabilities in the “toy dynamics” discussed in Section 1.8 is what distinguishes this choice of dynamics as a *classical stochastic theory*. A natural, and indeed inevitable, question arising in this context is *why*, at a fundamental level, the transition from “classical” to “quantum” is accompanied by a transition from \mathbb{R} to \mathbb{C} . Most attempts to address this question involve significant structural assumptions that narrow the focus to the specific properties of particular “state spaces.” For example, a 2001 paper of Caves, Fuchs, and Schack [CF02] explains why complex Hilbert spaces, but not real or quaternionic Hilbert spaces, support “reasonable behavior” for density operators in quantum theory.⁴⁵ A common “continuity” argument in favor of the use of complex numbers in quantum theory is that the unitary operations representing time evolution in this context should possess square roots, enabling the “time interval” under consideration to be subdivided.⁴⁶ Obviously, arguments of this type are generally irrelevant in the discrete causal setting, which does not involve the interpolative property.

⁴⁴Discrete causal theory raises the possibility of much more natural types of “hyperquantum theories,” defined by adding additional levels of hierarchy “above the level of kinematic schemes.” This topic is revisited in Section 7.10.

⁴⁵Another virtue of the paper [CF02] is that it supplies a spectrum of useful references on the subject.

⁴⁶For example, this argument is raised in the course of an illuminating general discussion of the subject on Scott Aaronson’s blog on quantum information theory.

However, discrete causal theory itself offers interesting insight into the quantum-theoretic role of \mathbb{C} , and also suggests that \mathbb{H} and \mathbb{O} remain worth considering in similar roles. The basic argument, which applies most naturally to the path summation approach to quantum theory in the background independent setting, is that one should not “artificially discriminate” among “evolutionary pathways” for classical histories, by assigning them weights of different *magnitudes* in path sums. This immediately focuses attention on *spheres*, or close analogues of spheres, as the natural target objects for the “phases” of these pathways, since spheres are the prototypical examples of “spaces whose elements all possess the same magnitude.” The desire to preserve the algebraic structure associated with “splicing together evolutionary pathways” then leads to the consideration of “multiplicative structures on sphere-like objects,” while the procedure of path summation requires “additive structures” on larger objects containing these “spheres.” Without attempting to be precise, these very general reflections lead immediately to the consideration of the 1-sphere S^1 , viewed as a subobject of \mathbb{C} , the 3-sphere S^3 , viewed as a subobject of \mathbb{H} , and the 7-sphere S^7 , viewed as a subobject of \mathbb{O} . Coincidentally or not, S^1 is the target of Feynman’s phase map in his original description of the path summation approach [FE48]. In the discrete setting, of course, it is more natural to consider subobjects of these spaces that are not real manifolds, or possibly abstract analogues of such spaces. These rather speculative ideas are revisited in Section 6.7, in the context of adapting the path summation approach to the discrete causal setting. Note that *none of this reasoning has anything to do with the “completeness” of such number systems, or suggests that they should serve as basic structural scaffolding for spacetime.*

The right-hand column of the table in Figure 2.9.2, which lists the order-theoretic structure of the number systems appearing in the table, reveals how the buildup from $\mathbb{N} - \{0\}$ to \mathbb{R} , and beyond, first obscures, then destroys, the order-theoretic significance of the term “number.” The notion of “quantity” suffers a similar fate, at least from a physical perspective, since cardinality provides much too coarse a description to support a useful notion of scale in continuum-based settings. Given the ongoing troubles of continuum-based theories, as well as the relatively barren physical role of the quaternions and octonions to date, it is worth considering the possibility that this progression of “more-and-more-complete” number systems is *simply not well-suited to describing fundamental physics*. In particular, as noted above, the general quantum-theoretic virtues of the field of complex numbers \mathbb{C} do not seem to depend on its “completeness.” For the simplest number systems, order-theoretic information, *together with* sufficiently-refined notions of quantity, are embodied in the same objects; for example, in the set of natural numbers \mathbb{N} , and in subsets of \mathbb{N} . Historically, this is one reason why the distinctions between ordinals and cardinals were not clearly recognized until the late nineteenth century. In the present context, \mathbb{N} stands in stark contrast to larger and more complicated number systems such as \mathbb{C} , which fail to preserve suitable notions of order and quantity, at least in physical settings. Fortunately, the discrete directed sets central to discrete causal theory, while they are much more complicated than the natural numbers \mathbb{N} , at least share with \mathbb{N} the important feature of supporting useful versions of *both* concepts. Among other important consequences, this is why metric recovery “works” in the discrete setting.

Physical objections to continuum-based theories. It is useful to briefly focus specific attention on a few of the many objections to the use of \mathbb{R} as a basic source of structure in fundamental physics. Some of these objections involve technical issues, while others are more conceptual in nature, or arise from the theory and practice of experimental science. The persistence of these problems suggests that the mainstream modern physics community should seriously consider devoting a more equitable proportion of its efforts to alternative approaches, which can potentially avoid these problems entirely.

1. *Divergence issues.* Many of the theoretical problems of continuum-based theories of fundamental physics involve divergences that arise precisely because certain quantities are permitted to become arbitrarily small. This is true, in particular, of general relativity and quantum field theory. It is especially difficult to construct continuum-based background independent quantum theories that avoid such divergences. These problems are deeply-rooted, and remain mostly intractable, despite several generations of intensive effort by the world's foremost physicists. There is no convincing evidence that they can be surmounted or circumvented without a significant change of structural paradigm. Similar problems are to be expected in almost any theory in which \mathbb{R} plays a substantial role.
2. *Lack of natural scale.* Real manifolds do not possess natural scale data. In particular, different coordinate charts on a real manifold X yield different "sizes" for a given subset. The only natural way to measure the "quantity" of a subset in this context is by its cardinality, which is much too coarse to provide a meaningful notion of scale in physical settings. For example, \mathbb{R}^n has the same cardinality as \mathbb{R} for any positive integer n . This means that scale data must be *supplied artificially* by the addition of auxiliary information, such as a metric. In the context of general relativity, this implies that metric recovery from causal structure is possible only up to a conformal factor, as established by Malament's theorem.
3. *Experimental discreteness.* The choice to discard continuum-based assumptions, even without prior theoretical justification, has proven strikingly successful historically. The prototypical example is Planck's solution of the blackbody radiation problem, which illustrates how divergence issues, arising from an irrelevant assumption of underlying continuum-based structure, can be cured by changing to a discrete paradigm. More generally, quantum theory has already replaced continua with discrete sets in a host of physical situations. Given this precedent, it seems imprudent to automatically retain continuum-based assumptions everywhere that experimental evidence has not already rooted them out.
4. *Discreteness arising from continuum-based assumptions.* Even continuum-based theories tend to predict important instances of discreteness in quantum-theoretic contexts. For example, in conventional quantum theory, certain operators on Hilbert spaces of functions over a real manifold may happen to possess discrete spectra of eigenvalues, leading to *derived* discreteness for the values of the associated observable quantities. At a more basic level, continuum-based approaches to quantum gravity and fundamental spacetime structure tend to *arrive at a form of fundamental discreteness*, via "quantization of spacetime." Most notably, loop

quantum gravity features “area” and “volume” operators that measure the “smallest meaningful units” of these observables. Thus, fundamental discreteness forces its way into the picture regardless of continuum-based assumptions.

5. *Discreteness via the philosophy of measurement.* It is impossible to directly establish, via measurement, the existence of a continuum of values of any observable quantity. In particular, one may always posit discrete structure at smaller scales, and such structure may be experimentally detectable, directly or indirectly. This realization, by itself, should not rule out continuum-based theories, but it should be considered as a mark against them, since it is awkward to invoke structure whose existence can never be established, even in principle. This view is closely related to a philosophical preference for background independence, since it treats as undesirable the practice of carrying along a “continuum background” in which discrete families of measurement values are taken to be “embedded.” As the theory of metric recovery illustrates, assuming the existence of such a “background” can lead to difficulties that are more than merely aesthetic, such as the loss of natural scale.

Historical and sociological objections to continuum-based theories. After demonstrating the existence of serious experimental, mathematical, and/or logical problems associated with an existing approach to fundamental physics, or a class of such approaches, it can be instructive to consider possible historical and sociological factors that might have contributed to these shortcomings. This is a risky exercise, due to the hazard of historical bias. In particular, it is tempting to view the development of scientific thought in a *teleological* sense, as “leading up to” the present state of scientific understanding, and this type of presumption often generates serious misconceptions. For example, popular “explanations” of Zeno’s paradoxes, which fortunately are ignored by serious physicists, mathematicians, and philosophers, focus on the largely irrelevant development of the rational and real number systems, boasting that “it is now known that an infinite number of terms can sum to a finite answer.” In fact, the original statements of these paradoxes are *physical* in nature, and essentially question whether or not spacetime possesses the interpolative property. As another example, Riemann seems to have been as ready to consider “discrete manifolds” in the 1850’s as he was to consider continua, but many physicists regard his work as “leading up to” general relativity. Despite these risks, it can still be useful to think about the development of science itself in terms of cause and effect. In this spirit, I mention the following factors as possibly contributing to the historical preeminence of continuum-based theories of fundamental physics.

1. *Pragmatism of continuum-based theories.* Until recently, continuum-based theories have been remarkably successful in providing solutions to the physical problems of greatest immediate interest, particularly in the context of applied science and engineering. In other words, these theories have prospered not because they seem likely to be *true*, but because they have been *useful*. A modern analogy is illuminating: during the construction of the standard model of particle theory, physicists were well-aware that the tools involved; in particular, background-dependent quantum field theories on Minkowski spacetime \mathbb{R}^{3+1} , were very

unlikely to address fundamental problems such as the unification of relativity and quantum theory.⁴⁷ However, these tools facilitated short-term progress, so they achieved temporary ascendancy. Failure to address deeper issues has largely squelched further progress over the last generation, and the lengthy run of success of continuum-based methods in fundamental physics since the time of Newton may well have ended with the completion of the standard model.

2. *Early ignorance of experimental discreteness.* Reasonable judgment regarding the adequacy of a structural paradigm is strongly influenced by the results of experiment and observation. In this regard, the last century has left continuum-based approaches to fundamental physics on far shakier ground than previously, due to recognition of numerous discrete phenomena via advances in microtechnology, which spurred the development of ordinary quantum theory. By this measure, early preeminence of continuum-based theories may be partly attributable to early ignorance of such discrete behavior at small scales.
3. *Early lack of structural alternatives to the continuum.* The continuum-based focus of early-modern physics was likely influenced by an absence of recognized mathematical alternatives to the real number system as a source of basic structure. Many structurally promising and physically suggestive alternatives arising in information theory, order theory, graph theory, and category theory were nonexistent or unrecognizable. Of course, the formal properties of \mathbb{R} were not explicitly understood during this period either, but this technical imprecision seems to have had little effect on the conceptual development of physics.
4. *Early lack of computational alternatives to continuum-based techniques.* The predominance of continuum-based methods may be partly attributed to a lack of computational alternatives to techniques from calculus, differential equations, and other areas of mathematics involving real analysis. Computational science was in its infancy when modern physics arose. Discrete models involving more than a few elements would have seemed computationally intractable even if they had been considered conceptually promising. For example, Riemann might have had difficulty actually *studying* examples of “discrete manifolds,” whatever his estimation of their fundamental merits. By contrast, real analysis is remarkably congenial to the computational limits of the unaided human brain, regardless of its ultimate physical relevance.

2.10 The Philosopher’s Peril

Scientific philosophy. A significant task of Chapters 3 and 4 of this book is to analyze the *physical plausibility* of various mathematical models for encoding discrete causal structure, and of axiomatic systems governing these models. I would prefer to regard much of this analysis as “appeals to the self-evident,” i.e., as common sense, but

⁴⁷Of course, there are important connections between Yang–Mills theory and general relativity in the context of loop quantum gravity.

objectively speaking, it belongs to the field of “scientific philosophy.” This field, of course, is one of the principal avenues whereby scientists and philosophers alike have made fools out of themselves since antiquity. Aristotle, for example, still receives what is probably an undeserved degree of ridicule for his unfortunate non-empirical conclusions regarding falling bodies. However, there is no shortage of other examples, both ancient and modern. Perhaps partly as a response to this, many modern physicists have taken the view that attempting to formulate novel approaches to fundamental physics based on “physical intuition” and “general principles” is a waste of time, and instead favor the more conservative approach of trying to first match experimental data before drawing any philosophical conclusions.⁴⁸ At an individual level, this strategy is pragmatic, since such efforts frequently achieve modest success, while more ambitious approaches usually fail. However, these “successful” models are often merely incremental updates of previous models, and suffer from the same obvious foundational issues as their progenitors. They also tend to exhibit what might be referred to as the “more-and-more disease,” in which a good idea from an established theory is extended beyond its scope of applicability. Historically, this has led to spurious consideration of more and more epicycles, particles, symmetries, dimensions, and so on, often accompanied by less and less progress in basic understanding.

Of course, *ignoring* experimental data is even worse; doing so pits the philosopher against the Almighty, a losing proposition.⁴⁹ Perhaps the sensible middle ground is to avoid choosing sides between science and scientific philosophy at all. Einstein certainly harbored a healthy respect for scientific philosophy, remarking that the deep physical principles underlying the natural world cannot be logically deduced, but must be reached by “*intuition, resting on sympathetic understanding of experience*” [EI34]. While mostly above reproach in his own right, Einstein is sometimes cited as a misleading exception to the “rule” that intuition is relatively useless compared to close grappling with experimental evidence. This rule seems valid, however, only if intuition is allowed to operate unconstrained by what is actually known. If the “sympathetic understanding of experience” is taken into consideration, then exceptions accumulate rapidly. For example, the Lagrangian and Hamiltonian formulations of mechanics and the path summation approach to quantum theory are all based on intuition involving deep general principles. Deep general ideas about spontaneous symmetry breaking, including the Higgs mechanism, were formulated years before acquiring a specific use in the Glashow–Weinberg–Salam electroweak theory. Non-abelian gauge theories were developed on basic structural grounds long before they were used to describe the electroweak and strong interactions. These examples sug-

⁴⁸Hawking has gone so far as to remark that “philosophy is dead.” While I disagree with this statement at face value, Hawking was primarily expressing a low opinion of the scientific competence of the philosophical community, *not* suggesting that philosophical issues themselves are vacuous. Ironically, Malament works as a philosopher.

⁴⁹Erdős remarked facetiously that life is a game against the “Supreme Fascist,” his fanciful conception of God. The human contestant can never score any points, but can keep the S.F.’s score as low as possible, by avoiding error.

gest that even intuition has its merits.⁵⁰ In any case, it may be healthy to balance the philosophical skepticism of the modern physics community, which has not enjoyed the privilege of seeing any revolutionary advances, with the outlook of a more fortunate generation. For example, Hermann Weyl expressed the following viewpoint in the introduction to his classic text *Space Time Matter* [WE52]:

And now, in our time, there has been unloosed a cataclysm which has swept away space, time, and matter, hitherto regarded as the firmest pillars of natural science, but only to make place for a view of things of wider scope, and entailing a deeper vision. This revolution was promoted essentially by the thought of one man, Albert Einstein... Philosophy, mathematics, and physics have each a share in the problems presented here... I shall only touch lightly on the philosophical implications, for the simple reason that in this direction nothing final has yet been reached, and that for my own part I am not in a position to give such answers to the epistemological questions involved as my conscience would allow me to uphold... As things stand today... the separate sciences... should follow in good faith the paths along which they are led by reasonable motives proper to their own peculiar methods and special limitations. The task of shedding philosophic light onto these questions is nonetheless an important one... This is the point at which the philosopher must exercise his discretion. If he keeps in view the boundary lines determined by the difficulties inherent in these problems, he may direct, but must not impede, the advance of sciences whose field of inquiry is confined to the domain of concrete objects. Nevertheless, I shall begin with a few reflections of a philosophical character... (pp. 2–3)

Weyl's carefully balanced perspective respects the necessity for caution in the practice of philosophy in the physical sciences, but never questions the legitimacy or importance of the discipline itself. Similar acknowledgment of the role of scientific philosophy has remained relatively mainstream throughout most of the history of modern science, despite current prejudices. This is not very surprising, because the subject involves questions of great scope and depth, whose difficulties remain as formidable, and whose consequences remain as important, as ever. Ultimately, the peril of squarely facing such difficulties must be accepted if these questions are ever to be adequately addressed.

Experimental challenges. In the last generation or so, an unfortunate practical limitation has grown increasingly prominent in fundamental physics: new experimental results in the realm of high energy particle theory have become almost prohibitively difficult and costly to obtain, thereby *forcing* philosophical substitutes to play a greater role in the way physics is done. For example, I have already mentioned the controversies surrounding the anthropic principle and the multiverse, in the context of string theory. Over much of the previous century, experiment decided the fate of many theoretical approaches within a few years of their inception. Today this is less-often true; a large proportion of theories that can be readily dismissed on experimental grounds are manifestly unworthy of attention in the first place, while most of the “interesting” theories are so difficult to test definitively that even decades-long research projects, and vast, internationally-funded engineering operations, cannot reliably decide their viability. This remains true, in particular, of string theory and

⁵⁰The mathematical reader will recognize here a slightly facetious reference to Gordan's reluctant endorsement of Hilbert's “theology.”

loop quantum gravity. Under these conditions, every possible source of insight into the conceptual integrity and technical viability of a theory is useful, including knowledge of where the theorist stands on philosophical grounds. Such information is far from definitive, since some great physicists have been suspect philosophers, while many outstanding philosophers have been terrible physicists. In this book, however, I wish to provide the reader with a viewpoint as conceptually comprehensive as possible. Hence, I devote the remainder of this section to the risky task of outlining some crucial conceptual and philosophical underpinnings of discrete causal theory, as I choose to approach it. The reader should be aware that important elements of this philosophy differ significantly from certain viewpoints associated with previously-existing versions of discrete causal theory.

Six basic principles. The following basic principles of scientific philosophy help to frame the overall viewpoint underlying the version of discrete causal theory developed in this book. In stating these principles, I make no attempt to be original; most of them fall well within the scientific mainstream. I also make no attempt to be definitive; it is possible to add or subtract a statement or two from this list without substantially altering the overall viewpoint. The remarks accompanying these statements provide illustrative examples, and indicate some of the places in the book where these principles are employed.

1. *Physics should seek not to prescribe what may be, but to describe and explain what is.* As explained in Section 2.1, the distinction between prescription and description is well-illustrated by comparing certain aspects of general relativity and discrete causal theory. General relativity treats the mathematical structure used to model classical spacetime as *prescribing* which pairs of events *may be* causally related. As noted in Section 1.3, this leads to “awkward counterfactual speculation” about actual events. Discrete causal theory, on the other hand, treats such structure as *describing* which pairs of events *are* causally related. Some might argue, with a degree of justification, that general relativity is merely *misinterpreted* along these lines,⁵¹ but this is indisputably the mainstream interpretation. Of course, some descriptions are more satisfying than others. An ideal description should *explain*, or “render intelligible,” the associated formalism, by showing it to faithfully represent clear fundamental principles.⁵² As outlined in the discussion of causality conditions in Section 2.7, the descriptive philosophy avoids meaningless inconsistencies, such as “time-travel paradoxes.” It also abstains from unjustified assumptions, such as the presumed steady state of the universe before Hubble’s observations. Looking ahead to Chapters 3 and 4, the axioms of *transitivity* (TR) and *interval finiteness* (IF), which feature prominently

⁵¹Here, I am thinking about influential scientists such as Rovelli [RO04], who offers a viewpoint about certain aspects of relativity that may give too much credit to the theory itself. I agree with much of the physical content of Rovelli’s viewpoint, without necessarily agreeing that Einstein’s theory itself suffices to adequately embody it.

⁵²Much has been written about whether or not physics should be *expected* to be intelligible, whatever the meaning of the term. In any case, it would be difficult to formulate a clearer fundamental principle than the causal metric hypothesis (CMH).

in the existing literature, are worrisomely prescriptive. For this reason and others, neither of these two axioms plays a role in the version of discrete causal theory developed in this book. This illustrates the fact that shifting to the discrete causal paradigm does not automatically cure prescriptive issues; the specific choice of axioms is also important. In a broader context, prescriptive issues are often related to a lack of perfect background independence, and more generally to the use of structures that are arbitrary, rather than structures possessing a universal property, as discussed in Section 1.6.

2. *Mathematics and physics are distinct; each informs the other.* Mathematical structures in physics should be chosen for their conceptual merits, not for their familiarity or convenience. This principle, which may be shortened to the phrase “*concept over convenience*,” recalls Gauss’ “*notions over notations*.” Historically, this practice has led not only to good physics, but also to interesting mathematics, while the mathematical community has returned the favor by introducing concepts and methods whose physical significance has only been appreciated much later. Care must be taken to distinguish between mathematical and physical properties. An unfortunate byproduct of the long-standing success of continuum-based theories in physics has been the automatic and unjustified attribution of certain *mathematical* properties of the continuum, such as the interpolative property and the completeness property discussed in Section 2.9, to physical spacetime. An important example of distinguishing between mathematical and physical properties in discrete causal theory is offered by the *independence convention (IC)*, introduced in Section 3.7. In this case, the two properties to be distinguished are mathematical *irreducibility* and physical *independence* of relations between pairs of elements in a directed or multidirected set.
3. *Basic structural concepts are crucial.* As mentioned in Section 2.9, it is instructive to consider the poverty of structural alternatives to continuum-based geometry available to physicists during the early 20th century, when relativity and quantum theory were developed. Many physically suggestive ideas from fields such as order theory, graph theory, information theory, computer science, category theory, algebra, and algebraic geometry, were not yet known. Even group theory faced a difficult reception, as evidenced by Wigner’s description of Schrödinger’s “*gruppenpest*.” The twenty-first century scientific community is much better equipped, at least on paper, to follow up Einstein’s intuition that physics is essentially structural in nature, brilliantly vindicated by general relativity, but largely unconsummated thereafter. Many of the structural ideas appearing in this book have roots in modern algebra, particularly in the work of Alexander Grothendieck. Especially important is Grothendieck’s *relative viewpoint (RV)*, formally introduced in Section 3.8, and applied in Chapters 4–7.
4. *Local and global properties must be properly distinguished.* The history of physics is littered with errors resulting from specious local-to-global reasoning and dubious extrapolation across scales. Often, such errors arise from failure to recognize the limitations of “obvious” observations, such as the apparent motionlessness of the earth, or the apparent flatness of spacetime in the vicinity of the earth, or the apparent possibility of assigning definite values of position and momentum simul-

taneously to macroscopic material bodies. Particularly troublesome are “local” conditions adopted without recognition of their global consequences. Looking ahead to Chapter 4, the axiom of interval finiteness (IF), sometimes mislabeled as “local finiteness” in the literature, prescribes *global* structure to an uncomfortable degree. This provides another illustration of why the specific set of axioms chosen for discrete causal theory is of crucial importance.

5. *The nature of experimentation has theoretical significance.* A specific instance of this principle, involving the uncomfortable status of continuum-based theories with regard to the philosophy of measurement, was mentioned in Section 2.9. More generally, besides attempting to explain specific experimental results, theorists should consider *general demands and prohibitions* associated with the experimental method. For example, in Section 1.3, I noted the unavoidable scientific role of directed relationships between experimental conditions and results. The nature of experimentation also favors axiomatizing local rather than global properties, since the latter may be experimentally inaccessible. For example, in the context of relativity, it is reasonable to assume that classical spacetime is four-dimensional, since dimension is defined locally, but it is unreasonable to assume a specific global topology. Of course, conclusions about large-scale topology could conceivably be *derived* on dynamical grounds, or inferred from unlikely observational scenarios, such as “circles in the sky.”
6. *Censor the fatal, not the merely unexpected.* A reasonable facet of theory-building is to impose conditions “censoring” properties that are so qualitatively contrary to observation that any theory exhibiting them is immediately discredited. More succinctly, it is reasonable to “ignore the irrelevant.” This is the rationale behind the discussion of narrowing the focus of the classical causal metric hypothesis (CCMH) to “physically relevant directed sets,” in Section 2.2. More generally, this principle defines the boundary of the prohibition against “prescription” mentioned above, by permitting the *proscription* of “fatal phenomena.” However, this idea must be applied with great care, due to the limitations of human judgment and imagination. Planck’s approach to black-body radiation, eliminating all but a discrete set of emission frequencies, to avoid the fatal ultraviolet catastrophe, is an example of justified censorship. However, Einstein’s fixing of the “cosmological constant,” to achieve his *expectation* of a steady-state universe, is not. In the context of discrete causal theory, the axioms of *transitivity* (TR) and *interval finiteness* (IF), already mentioned above, censor *nonfatal* phenomena in problematic ways; the former by ignoring distinctions among certain modes of influence between pairs of events, and the latter by drastically constraining the global structure of classical spacetime.

Ten qualitative assumptions of discrete causal theory. The six principles listed above are very general, and provide only a partial overview of a broad approach to doing science. They *inform* the developments described in this book in important ways, but do not come close to determining them. Hence, it is useful to gather together some more-specific assumptions underlying the version of discrete causal theory developed here. Many of these assumptions have already been mentioned

and applied, at least implicitly, in previous sections, but they have not yet appeared explicitly in one place. Among other advantages, this listing provides the opportunity to revisit some of the conceptual topics discussed in the questions and answers at the end of Chapter 1, with the benefit of additional information from the present chapter.

1. *There is no physical continuum.* In Section 2.9, I discussed some of the general shortcomings of physical models that rely on the real number continuum \mathbb{R} as a source of basic structure. Discrete causal theory treats these objections as sufficient to completely rule out the use of \mathbb{R} in this role.
2. *The physical universe is basically discrete.* Having dispensed with \mathbb{R} , along with derivative structures such as real manifolds, it is necessary to offer an alternative structural paradigm for fundamental physics. Discrete structures seem to furnish more-natural and more-promising physical models. It is important to emphasize that discreteness is *not* “the exclusive physical alternative” to the continuum; most of the structural paradigms one could choose to explore are neither continuous nor discrete. However, there are good reasons why most modern approaches to fundamental physics rely heavily on one or both of these extremes: they possess special properties that are either physically suggestive, or amenable to mathematical analysis, or both. From the viewpoint of discrete causal theory, the conceptual advantages of discrete models far outweigh the mathematical convenience of continuum-based models.
3. *Physics is about cause and effect.* This statement is, of course, a paraphrase of the causal metric hypothesis (CMH), which is the subject of the present chapter, and one of the main themes of the book. The first two chapters have presented a number of different versions, paraphrases, and shades of meaning of this hypothesis. A crucial part of the picture, not yet discussed in detail, is the *quantum causal metric hypothesis* (QCMH), which requires additional theoretical background before it can be stated in a precise manner. The necessary developments are carried out in Chapters 5–7.
4. *Classical spacetime may be modeled in terms of directed sets.* This statement is a paraphrase of the classical causal metric hypothesis (CCMH). As explained in Section 2.2, discrete classical causal structure may be modeled, at the local level, in terms of directed relationships between pairs of elements, and a directed set is merely a collection of such related pairs, considered as a single object. One advantage of describing causal structure in this way is that it avoids prescribing dubious global properties, such as transitivity (TR). More-detailed discussion of axiomatic systems for directed sets appears in Chapters 3 and 4.
5. *Quantum spacetime may be modeled in terms of multidirected sets.* This is the essence of the quantum causal metric hypothesis, although the details are postponed until Chapter 7. Multidirected sets are natural generalizations of directed sets, in which a given pair of elements may have multiple relations between them in either or both directions. The elements in discrete quantum causal theory represent classical histories, and the corresponding relations represent relationships between classical histories, i.e., co-relative histories. Similarly, the multidirected

sets arising in this context represent kinematic schemes. The reason why multi-directed structure, and not merely directed structure, is necessary in this context, is because of a subtle technical property of directed sets that permits the existence of multiple distinct co-relative histories between a given pair of classical histories.

6. *Classical histories are generally nontransitive.* The conceptual basis of this assertion is merely that direct and indirect relationships are physically different; the details are explained in Chapter 3, particularly in Section 3.9. Using a nontransitive binary relation, direct causation may be modeled by an individual relation between a pair of events, while indirect causation may be modeled by chains of relations. In general relativity, influence is taken to flow along causal curves, which implies that *every* instance of causation is indirect, due to the interpolative property of \mathbb{R} . However, in the discrete setting, both direct and indirect influences are possible, and it is necessary to use models that enable distinction between the two. Following this reasoning, the possibility of direct causation is an *essentially new feature*, introduced by exchanging continuum-based models for discrete ones.
7. *Classical histories are locally finite, i.e., star finite (SF), but not necessarily interval finite (IF).* The motivation for imposing a local finiteness condition on discrete causal structure is, roughly speaking, that one expects individual elements to possess a “finite size” in the discrete context. Hence, one faces convergence issues if an infinite number of elements are permitted to coexist in a “local region.” The subject of local behavior in discrete causal theory is examined in much more detail in Chapter 4, but I briefly elaborate on the need for a local finiteness condition here. The metric recovery results discussed earlier in the present chapter imply that if natural scale data can somehow be derived from causal structure, then one can recover the apparent geometric properties of relativistic spacetime at ordinary scales. The obvious way to proceed in the discrete context is by “using local combinatorial data to determine scale,” generalizing the causal set prescription. To avoid physical pathologies, such as the instantaneous expansion of a minimal region of spacetime into an infinite volume, it is natural to impose the condition that “each element is directly related to only a finite number of other elements.” This local finiteness condition, called *star finiteness*, is formally introduced in Section 4.4. Unfortunately, the term “local finiteness” is sometimes used in the literature to denote the very different condition of *interval finiteness*, which is not a local condition at all, and which permits the very type of physical pathologies described above.
8. *The relative viewpoint is indispensable.* As explained in Section 1.5, Grothendieck’s relative viewpoint (RV) embodies the philosophy that objects should not be analyzed in isolation, but should be studied along with their natural relationships. At the classical level, this viewpoint leads to the definition of the *relation space* $\mathcal{R}(D)$ over a directed set D , viewed as a discrete causal classical history. Relation space was briefly introduced in Section 1.5, and is studied in detail in Chapter 5. At the quantum level, the relative viewpoint leads to the theory of *co-relative histories* and *kinematic schemes*, developed in Chapters 6 and 7.

9. “Classical” and “quantum” may be understood as levels of structural hierarchy. As described in the first few pages of the book, discrete causal theory exhibits an attractive self-similarity, called *iteration of structure* (IS), in which quantum structure naturally occupies a level of mathematical hierarchy above that of classical structure. A hint of this relationship is evident even in the ordinary Hilbert-space approach to quantum theory, which shifts the focus from individual states, and relationships among them, to spaces of states, and operators on these spaces. Category theory provides a more general analogy, elaborated in a striking manner by Christopher Isham [IS05]: “quantization” corresponds roughly to passage from “elements and relations” to “objects and morphisms.” What is special, and perhaps unique, about discrete causal theory, in this context, is that its “higher-level quantum objects” possess essentially the same type of structure as its “lower-level classical objects.” A concise way to paraphrase this viewpoint is to say that “classical physics is about relationships between pairs of events; quantum physics is about relationships between pairs of histories.”
10. *Quantum dynamics arises from generalized path summation.* The import of this assertion is that one particular approach to quantum theory; namely, Feynman’s *path summation approach*, is sufficiently general to apply to discrete causal theory, for which other popular approaches are inadequate. This is because most approaches to quantum theory take for granted a great deal of structure arising from the properties of the real and complex numbers, which is generally unavailable in the discrete causal setting. For example, ordinary quantum theory and quantum field theory begin with Hilbert spaces of complex-valued functions over real manifolds, which depend on constructions that discrete causal theory treats as emergent.⁵³ By contrast, the path summation approach may be abstracted to apply to any situation involving families of directed relationships between pairs of classical histories.

References

- [HA76] S. W. Hawking, A. R. King, and P. J. McCarthy. *A new topology for curved space-time which incorporates the causal, differential, and conformal structures.* Journal of Mathematical Physics, **17**, 2, pp. 174–181, 1976.
- [MA77] David B. Malament. *The class of continuous timelike curves determines the topology of spacetime.* Journal of Mathematical Physics, **18**, 7, pp. 1399–1404, 1977.
- [WH98] John Archibald Wheeler and Kenneth W. Ford. *Geons, Black Holes, and Quantum Foam: A Life in Physics.* W. W. Norton and Company, New York, 1998.
- [IS05] Christopher Isham. *Quantising on a Category.* Foundations of Physics, **35**, 2, pp. 271–297, 2005. arXiv preprint: <http://arxiv.org/pdf/quant-ph/0401175v1.pdf>.
- [SO12] Rafael Sorkin. *Toward a Fundamental Theorem of Quantal Measure Theory.* Mathematical Structures in Computer Science, **22**, 05 (special issue), pp. 816–852, 2012. arXiv preprint: <http://arxiv.org/pdf/1104.0997v2.pdf>.

⁵³ An alternative, of course, is to abstract, elevate and generalize this algebraic structure; this is what is done in Connes’ *noncommutative geometry*.

- [CA15] S. Carlip. *Dimensional reduction in causal set gravity*. Classical and Quantum Gravity, **32**, 23, 232001, 2015. arXiv preprint: <http://arxiv.org/pdf/1506.08775v3.pdf>.
- [RO04] Carlo Rovelli. *Quantum Gravity*. Cambridge Monographs on Mathematical Physics. Cambridge University Press, 2004.
- [ZE64] E. C. Zeeman. *Causality Implies the Lorentz Group*. Journal of Mathematical Physics, **5**, 4, pp. 490–493, 1964.
- [CA04] Sean Carroll. *Spacetime and Geometry: An Introduction to General Relativity* Addison Wesley, 2004.
- [HE73] S. W. Hawking and G. F. R. Ellis. *The large scale structure of space-time*. Cambridge Monographs on Mathematical Physics. Cambridge University Press, 1973.
- [MA99] Juan Maldacena. *The Large N Limit of Superconformal Field Theories and Supergravity*. International Journal of Theoretical Physics, **38**, 4, pp. 1113–1133, 1999.
- [EMM12] George F. R. Ellis, Roy Maartens, and Malcolm A. H. MacCallum. *Relativistic Cosmology*. Cambridge University Press, 2012.
- [BLMS88] Luca Bombelli, Joohan Lee, David Meyer, and Rafael Sorkin. *Bombelli et al. Reply to Comment on “Space-Time as a Causal Set.”* Physical Review Letters, **60**, 7, pp. 656, 1988.
- [WA84] Robert M. Wald *General Relativity*. University of Chicago Press, 1984.
- [BA12] Julian Barbour. *Shape Dynamics: an Introduction*. Quantum Field Theory and Gravity: Conceptual and Mathematical Advances in the Search for a Unified Framework, Edited by Felix Finster, Olaf Müller, Marc Hardmann, Jürgen Tolksdorf, and Eberhard Zeidler. arXiv preprint: <http://arxiv.org/pdf/1105.0183v1.pdf>.
- [PE10] Roger Penrose. *Cycles of Time*. Vintage Books, New York, 2010.
- [TH07] Thomas Thiemann. *Modern Canonical Quantum General Relativity*. Cambridge Monographs on Mathematical Physics. Cambridge University Press, 2007.
- [BS07] Antonio N. Bernal and Miguel Sanchez. *Globally hyperbolic spacetimes can be defined as “causal” instead of “strongly causal”*. Classical and Quantum Gravity, **24**, 3, 745, 2007. arXiv preprint: <http://arxiv.org/pdf/gr-qc/0611138v1.pdf>.
- [BZ15] Ali Bleybel and Abdallah Zaiour. *A general theorem on temporal foliation of causal sets*. Preprint, 2015. arXiv preprint: <http://arxiv.org/pdf/1508.01052v1.pdf>.
- [MRS06] Seth A. Major, David Rideout, and Sumati Surya. *Spatial Hypersurfaces in Causal Set Cosmology*. Classical and Quantum Gravity, **23**, 14, pp. 4743–4751, 2006. arXiv preprint: <http://arxiv.org/pdf/gr-qc/0506133v2.pdf>.
- [BO89] Luca Bombelli and David Meyer. *Origin of Lorentzian geometry*. Physics Letters A, **141**, 5-6, pp. 226–228, 1989.
- [MP06] Keye Martin and Prakash Panangaden. *A Domain of Spacetime Intervals in General Relativity*. Communications in Mathematical Physics, **267**, 3, pp. 563–586, 2006.
- [PS11] Onkar Parrikar and Sumati Surya. *Causal topology in future and past distinguishing spacetimes*. Classical and Quantum Gravity, **28**, 15, 155020, 2011.
- [BHS09] Luca Bombelli, Joe Henson, Rafael Sorkin. *Discreteness without symmetry breaking: a theorem*. Modern Physics Letters A, **24**, 32, pp. 2579–2587, 2009. arXiv preprint: <http://arxiv.org/pdf/gr-qc/0605006v1.pdf>
- [CSS16] Hua Chen, Naoki Sasakura, and Yuki Sato. *Emergent classical geometries on boundaries of randomly connected tensor networks*. Preprint, 2016. arXiv preprint: <http://arxiv.org/abs/1601.04232>
- [CF02] Carlton M. Caves, Christopher A. Fuchs, and Rüdiger Schack. *Unknown Quantum States: The Quantum de Finetti Representation*. Journal of Mathematical Physics, **43**, 9, pp. 4537–4559, 2002. arXiv preprint: <http://arxiv.org/pdf/quant-ph/0104088v1.pdf>
- [FE48] Richard Feynman. *Space-Time Approach to Non-Relativistic Quantum Mechanics*. Reviews of Modern Physics, **20**, 2, pp. 367–387, 1948.
- [EI34] Albert Einstein. *Essays in Science*. Philosophical Library, New York, 1934.
- [WE52] Hermann Weyl. *Space Time Matter*. Dover, 1952.

Discrete Causal Theory

Emergent Spacetime and the Causal Metric Hypothesis

Dribus, B.F.

2017, XXX, 558 p. 154 illus., Hardcover

ISBN: 978-3-319-50081-2